Journal of Biomolecular NMR, **25:** 41–53, 2003.
KLUWER/ESCOM
© 2003 *Kluwer Academic Publishers. Printed in the Netherlands.*

Prediction algorithm for amino acid types with their secondary structure in proteins (PLATON) using chemical shifts

D. Labudde, D. Leitner, M. Krüger & H. Oschkinat*
Forschungsinstitut für Molekulare Pharmakologie, Robert-Rössle-Str. 10, 13125 Berlin, Germany

Received 7 August 2002; Accepted 25 September 2002

Key words: chemical shift pattern, PLATON, prediction algorithm, protein NMR, secondary structure, statistics, PLATON

Abstract

The algorithm PLATON is able to assign sets of chemical shifts derived from a single residue to amino acid types with its secondary structure (amino acid species). A subsequent ranking procedure using optionally two different penalty functions yields predictions for possible amino acid species for the given set of chemical shifts. This was demonstrated in the case of the α-spectrin SH3 domain and applied to 9 further protein data sets taken from the BioMagRes database. A database consisting of reference chemical shift patterns (reference CSPs) was generated from assigned chemical shifts of proteins with known 3D-structure. This reference CSP database is used in our approach for extracting distributions of amino acid types with their most likely secondary structure elements (namely α-helix, β-sheet, and coil) for single amino acids by comparison with query CSPs. Results obtained for the 10 investigated proteins indicates that the percentage of correct amino acid species in the first three positions in the ranking list, ranges from 71.4% to 93.2% for the more favorable penalty function. Where only the top result of the ranking list for these 10 proteins is considered, 36.5% to 83.1% of the amino acid species are correctly predicted. The main advantage of our approach, over other methods that rely on average chemical shift values is the ability to increase database content by incorporating newly derived CSPs, and therefore to improve PLATON's performance over time.

Introduction

Protein structure determination by NMR is now a routine process in structural biology and structural genomics (Heinemann et al., 2000). Despite recent improvements of the procedure, however, it is still time consuming to determine a high resolution protein structure by NMR. Recent initiatives for high throughput structure determination use automated assignment routines (Moseley and Montelione, 1999) or alternatively selective labeling strategies (Medek et al., 2000). All automated assignment procedures rely on routines which evaluate their output lists (Moseley et al., 2001; Leutner et al., 1998; Bartels et al., 1997).

In this context, chemical shift analysis has been proven to be useful for validating assignments (Grzesiek and Bax, 1993), predicting secondary structure (Wishart and Sykes, 1994) and for providing additional constraints to be used in structure calculations (Wishart and Case, 2001). Examples are the programs PROTYP (Grzesiek and Bax, 1993) which identifies amino acid types based on Cα and Cβ chemical shifts, CSI which predicts secondary structures in proteins using assigned optional combinations of ^{13}C and ^{1}H chemical shifts (Wishart and Sykes, 1994), and TALOS (Cornilescu et al., 1999) which predicts Ψ and Φ backbone angles from assigned ^{13}C, ^{1}H, and ^{15}N chemical shifts. It has been shown that especially the use of heteronuclear chemical shifts significantly improves the reliability of the proposed results either for the assignment of amino acid types or the secondary structure prediction in proteins. However, other authors (Pons and Delsuc, 1999; Huang et al., 1997) demonstrated that it is possible to predict either amino

*To whom correspondence should be addressed. E-mail: oschkinat@fmp-berlin.de acid types solely on the basis of ^{1}H chemical shifts or amino acid types with their secondary structure elements employing ^{1}H chemical shifts and coupling constants. The authors are using neural networks. Both methods rely on separating complete amino acid spin systems by recording ^{15}N correlated spectra and neither method distinguishes all types of amino acids, only classes of amino acids. Crucial to all methods is the correct referencing of the chemical shifts (Wishart and Nip, 1998).

The aim of this work was the development of an algorithm which is able to predict both the amino acid type and the secondary structure element, based on a set of chemical shift values for the spin system of the respective amino acid as obtained from standard triple resonance experiments (Kay, 1997). We denote from now on amino acid type with the involved secondary structure element as amino acid species. A database consisting of specific chemical shift pattern (reference CSPs) was generated from assigned chemical shifts of proteins with known 3D-structure. These reference CSPs are the basis in our approach for the prediction of amino acid species. The program PLATON compares query CSPs of unassigned chemical shifts to the reference CSP database to predict the amino acid species and distinguishes between three secondary structure elements, namely α-helix, β-sheet, and coil. The input for the database were 51 protein structures.

We have shown that it is possible to predict amino acid species solely on the basis of CSPs using unassigned sets of chemical shifts of amino acid spin systems as an input.

Prediction algorithm of amino acid type and secondary structure: PLATON

Pattern definition of the reference CSP database

A chemical shift pattern (CSP), which is a vector of Booleans describing relative positions of chemical shifts, is defined by an optional combination of chemical shifts. The starting point for the definition of the CSP is the creation of an N-dimensional chemical shift space. N is determined by the kind of nuclei for which chemical shifts are available in the databases, for example Cα, Cβ, CO, and Hα, or subgroups of those. An example is given in Figure 1 for Cα, Cβ, and CO chemical shifts. The CSP contains a '+' or '−' as elements depending on the position of the investigated chemical shift with respect to a reference value, for all nuclei considered. The positions of the '+' and '−' are defined by the axis of the chemical shift space, for example CSP (Cα, Cβ, CO) = + − +. In Figure 1, the reference value is in the center of this three-dimensional chemical shift space (central red point). The chemical shift values of an amino acid are compared to this point. If the value is bigger a '+' is assigned, and analogously a '−' is assigned if the value is smaller. Hence for all dots in the red quadrant the CSP (Cα, Cβ, CO) = + + + is obtained. The chemical shift space can be further subdivided by introducing again reference points into the two halves of each dimension to allow for a distinction of otherwise identical CSPs. This is shown using the example of the two blue dots which have the same subpattern after the first iteration (CSP (Cα, Cβ, CO) = + + +). The new reference value (upper red point) defines another coordinate system in the upper right quadrant. Practically, the second and higher order reference points are chosen according to a statistical analysis of all amino acid species having the same three digit CSPs in the previous coordinate system.

Now each dot can be distinguished by the second set of coordinates (indexed with an apostrophe), yielding the vector CSPs, CSP_1 (Cα, Cβ, CO, Cα′, Cβ′, CO′) = ++++++ and CSP_2 (Cα, Cβ, CO, Cα′, Cβ′, CO′) = + + + − −−. The length of the CSP depends thus on the number of chemical shift space dimensions and on the number of created subspaces. This approach was chosen to take into account the continuity of the chemical shift space and the non-Gaussian distribution of chemical shifts.

Reference CSP database construction

Data obtained from 51 proteins were used to create a reference CSP database (see below). 15 data sets were taken from the TALOS database. A Tcl/Tk-script was written to extract chemical shifts, along with the protein identification, the residue number, and the PDB ID, from the BioMagResBank (BMRB) database (Seavey et al., 1991; http://www.bmrb.wisc.edu). All proteins with paramagnetic center and protein complexes were excluded. A second criterion was the availability of all backbone and Cβ chemical shifts. By applying these selection criteria a further 36 proteins were obtained. Hence PLATON uses Booleans to construct CSPs and therefore does not use the chemical shift values in the further course of the prediction, it is tolerant of slight variations in the reference. Our data were selected with no particular preference of a

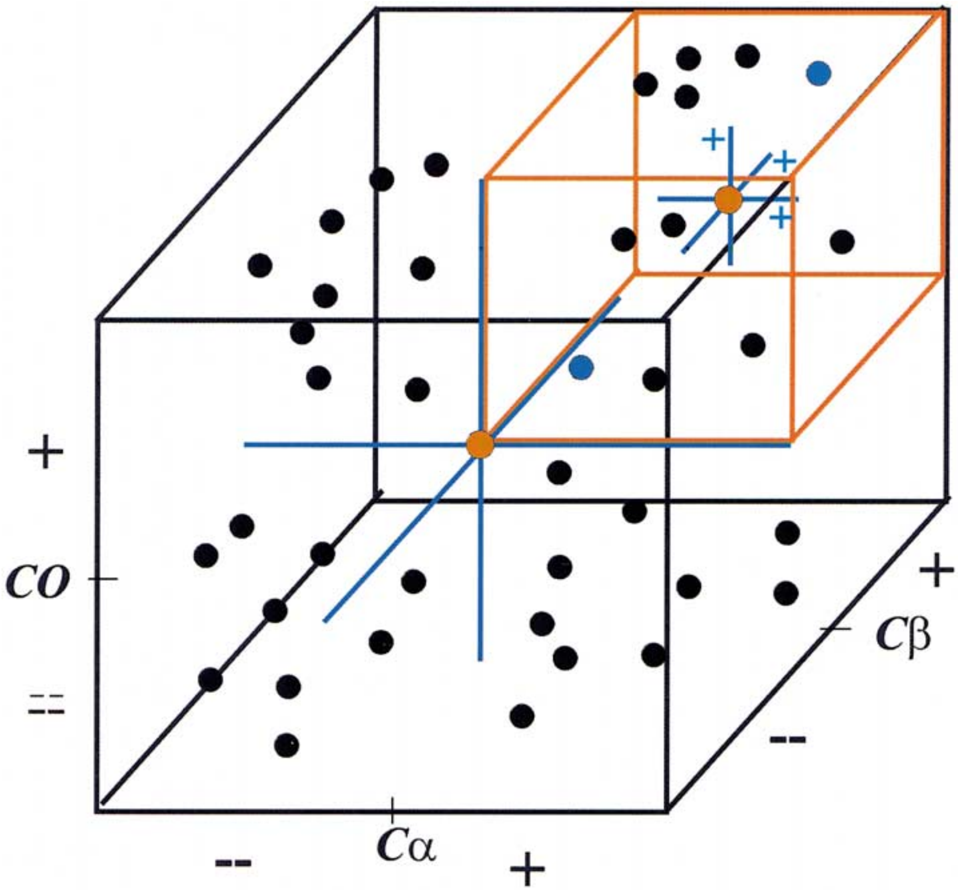

Figure 1. Illustration of the creation of CSPs in a N-dimensional chemical shift space (black frame). In this example Cα, Cβ, and CO chemical shifts were taken into account. The black dots indicate the chemical shifts of an individual amino acid. The red dots represent calculated center of masses which serve as reference values for the CSP building. The red frame is obtained after the first subdivision of the chemical shift space.

referencing system. However, most data (around 90%) are referenced to DSS (81%) or TSP (9%). Analysis of the pdb header file from the PDB database (http://www.pdb.org; Bernstein et al., 1977) revealed the secondary structure elements for each amino acid.

We used proteins belonging to a variety of protein classes according to SCOP (Structural Classification of Proteins; Murzin et al., 1995), whereby six SCOP structural classes are now represented in the reference CSP database: all-α, all-β, α/β, alpha and beta, membrane and cell surface proteins and peptides, and small proteins. The CO, Cα, Cβ, Hα chemical shifts and the corresponding secondary structure elements of 5896 amino acids served thus as input for the generation of the reference CSP database. In Figure 2, the relative frequencies of occurrence of all amino acids (w/o glycines) with the three used secondary structure elements are shown.

The Chou–Fasman (Chou and Fasman, 1974; Kyngäs and Valjakka, 1998; Fasman, 1989) parameters for the 51 proteins were calculated in order to control if our data selection reflects a representative protein structure space. They show the same trend (result not shown) for the frequency of occurrence of an amino acid type in α-helices and β-sheets, as predicted by Chou–Fasman (Chou and Fasman, 1974).

In Figure 3, a subset of the complete reference CSP database containing chemical shifts of 5896 amino acids is shown. This database is constructed of Cα, Cβ, CO, and Hα chemical shifts, using one subspace, yielding 256 eight digit reference CSPs. However, 7 reference CSPs did not correspond to a chemical shift constellation in the existing database of 5896 amino acids. They were not further considered. Due to the fact that each amino acid species possesses different chemical shift values in the N-dimensional chemical

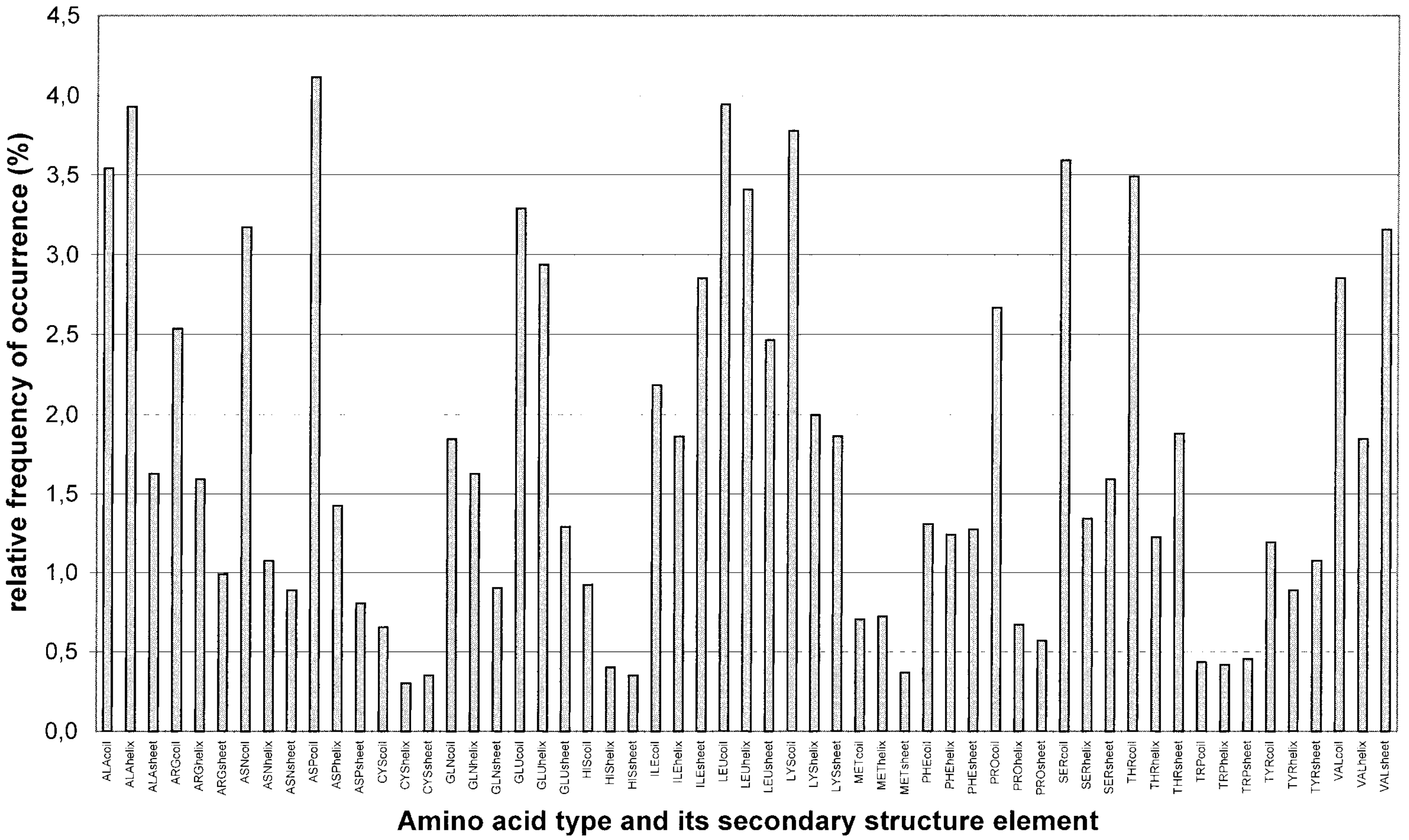

Figure 2. Relative frequency of occurrence of all amino acids (without glycine) with its secondary structure elements. The secondary structure elements are divided into three classes, namely α-helix, β-sheet, and coil (all other elements). The data were obtained from 51 proteins whose 3D structures are known.

No.	CSP	ALA			ARG			ASN			ASP			CYS			TYR			VAL			Sum2	No2	%
		C	H	S	C	H	S	C	H	S	C	H	S	C	H	S	C	H	S	C	H	S			
1	--------	4	3	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	13	7	12%
2	--------	0	0	0	1	0	0	5	0	1	0	0	0	0	0	0	0	0	0	0	0	0	14	8	14%
3	--------	6	2	3	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	23	9	16%
4	--------	0	1	0	4	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	13	10	18%
5	--------	7	7	3	2	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	26	10	18%
6	---+----	0	0	0	0	0	0	1	0	0	2	0	0	0	0	0	0	0	0	0	0	0	8	4	7%
7	--+-----	5	0	5	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	32	11	19%
8	-+------	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	10	6	11%
9	+-------	22	5	4	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	36	6	11%
10	------++	0	0	0	1	0	4	11	1	2	2	1	0	0	0	0	0	0	0	0	0	0	35	15	26%
11	-----+-+	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	8	7	12%
12	----+--+	0	0	0	0	0	0	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5	1	2%
13	---+---+	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	2%
14	--+----+	0	0	0	5	0	3	3	0	1	0	0	0	1	0	0	0	0	0	0	0	0	35	14	25%
15	-+-----+	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	8	6	11%
16	+------+	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4	4	7%
17	-----++-	0	0	0	2	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	10	6	11%
18	----+-+-	9	2	7	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	36	12	21%
19	---+--+-	0	0	0	0	0	0	1	0	1	2	0	1	0	0	0	0	0	0	0	0	0	10	7	12%
20	--+---+-	0	0	4	0	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	18	11	19%
21	-+----+-	0	0	0	0	4	0	0	0	0	0	0	0	1	1	0	0	0	0	4	0	0	20	10	18%
22	+-----+-	29	1	1	7	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	58	9	16%
23	----++--	0	0	0	6	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	29	8	14%
24	---+-+--	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	9	5	9%
233	++++-+-+	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5	2	4%
234	+++++--+	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	2%
235	-++++++-	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	19	2	4%
236	+-+++++-	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	1	0	0	0	0	0	4	3	5%
237	++-++++-	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	10	0	0	0	0	25	5	9%
238	+++-+++-	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	4	4	7%
239	++++-++-	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	8	3	5%
240	+++++-+-	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	3	3	5%
241	++++++--	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4	4	7%
242	-+++++++	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	16	2	4%
243	+-++++++	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	6	4	7%
244	++-+++++	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	22	4	7%
245	+++-++++	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	2%
246	++++-+++	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4	2	4%
247	+++++-++	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	2%
248	++++++-+	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	6	3	5%
249	++++++++	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4	2	4%
sum1:		211	232	96	150	95	58	186	64	53	239	85	48	39	18	21	69	52	64	169	110	183	5896		
number1:		19	16	16	52	22	31	34	30	19	38	20	19	28	16	16	39	27	28	44	17	37			
%		7%	6%	6%	20%	9%	12%	13%	12%	7%	15%	8%	7%	11%	6%	6%	15%	11%	11%	17%	7%	14%			

Figure 3. A subset of the complete reference CSP database created from Cα, Cβ, CO, and Hα chemical shifts is shown. The complete database contains reference CSPs from 5896 amino acids. The amino acids types are subdivided into three classes upon their secondary structure elements (C = coil, H = α helix, and S = β sheet). Sum 1 is the total number of the different amino acid species, for example ALAcoil occurs 211 times in the database. Number 1 is the number of occurring reference CSPs for a particular amino acid species contained in the database, for example for ALAcoil 19 different reference CSPs exist. This corresponds to 7% of all possible reference CSPs. Sum 2 is the total number of occurrence of each particular reference CSP. Number 2 is the number of amino acid species characterized by each particular reference CSP. The content of each reference CSP is a distribution of occurring amino acid species. The resulting percentage indicates how often a certain reference CSP occurs for different amino acid species.

shift space, several reference CSPs might be connected with one amino acid species. On the other hand, an amino acid species occurs predominantly in a defined region of the chemical shift space. Due to the small dispersion of the chemical shift values of different amino acid species, one reference CSP might represent several amino acid species.

For example, reference CSP 9 (+ − − − − − − −) contains six different entries (four are shown in Figure 3) which are only 11% of all possible amino acid species. These amino acid species with frequency of occurrence in the database are: ALAcoil (22), ALAhelix (5), ALAsheet (4), ARGsheet (1), GLNcoil (3), and GLNhelix (1) (the latter two are not shown in Figure 3). In the following, it will be used that CSPs of the same kind of unassigned amino acid species correspond to the one of highest occurrence, in this case ALAcoil. The justification for this is seen in the fact that all amino acid species show up in a limited amount of reference CSPs, for example ALAcoil is defined by only 19 different reference CSPs of 256 possible ones. The significance of the procedure should increase when a larger number of nuclei is taken into account. For further differentiation, the CSP can be extended by subdivision of the chemical shift space into smaller partitions. Each subdivision extends the CSP with an additional subpattern which would be in the example above four further digits per subdivision step (iteration).

PLATON algorithm

The first step is to execute a module that generates a reference CSP database with different chemical shift spaces. Chemical shift values extracted from public NMR databases serve as input for the reference CSP database. Secondly, a module compares the reference CSPs with CSPs of an investigated protein (query CSPs). Query CSPs are constructed in the same manner as reference CSPs. In Figure 4, a flow chart schematically illustrates the PLATON algorithm. The outlined steps of the algorithm are describing in more detail in the following paragraphs.

For investigated query CSPs, the rows in Figure 3 that correspond to the respective reference CSP yield an amino acid species distribution that may be interpreted further using penalty functions. We tested two penalty functions which are derived from different statistical assumptions.

Firstly, calculating the relative frequencies of occurrence of the amino acid species in the distributions leads to penalty function 1. This function is the ratio of the number of a certain amino acid species for a particular CSP and the total number of this amino acid species in the database (for example 13.7% for ALAcoil with CSP 22 (+ − − − − − + −). Secondly, the sum of the normalized variances of all average chemical shift values from the amino acid species used in PLATON represents directly the penalty value for each entry of the amino acid species distribution. The penalty options are used to associate each possibility with a probability-like factor, allowing for a ranking of the distributions.

Test run on α-spectrin SH3 domain

The procedure was tested using the chemical shifts of the α-spectrin SH3 domain while the database was constructed without these data. In addition, the parametrization was refined and the ability of the algorithm to incorporate new CSPs was examined.

The α-spectrin SH3 domain consists of 62 amino acids. Both the X-ray and the NMR structures are known (Musacchio et al., 1994; Blanco et al., 1997). At first, query CSPs with CO, Cα, Cβ and Hα chemical shifts were constructed of those 57 amino acids for which the respective chemical shift data were available. The α-spectrin SH3 domain contains three glycines which were not considered. Furthermore, Ser19 and Met1 were excluded due to the fact that the CO chemical shifts were not assigned by Blanco et al. For 52 out the 57, the correct amino acid species were contained in the distributions that resulted from the inspection of the reference CSP database. The queries for Tyr15, Trp41, Trp42, Pro54, and Lys60 yielded results without the respective amino acid species.

The distributions were alternatively ranked using the penalty functions (see PLATON algorithm). The results of the α-spectrin SH3 domain after applying the different penalty functions are shown in Figure 5. Circles represent the output of PLATON penalized with the relative frequencies of occurrence of the CSPs (penalty function 1), while squares indicate the application of the second penalty function (sum of the normalized variances). For 32 residues the two different penalty functions lead to significantly different values in the two ranked distributions.

Upon application of the two penalty functions the two resulting ranking lists were compared. In both lists 20 amino acid species are listed on the same ranking positions. Out of these 20, 14 amino acid species are predicted in position one. This means that highly re-

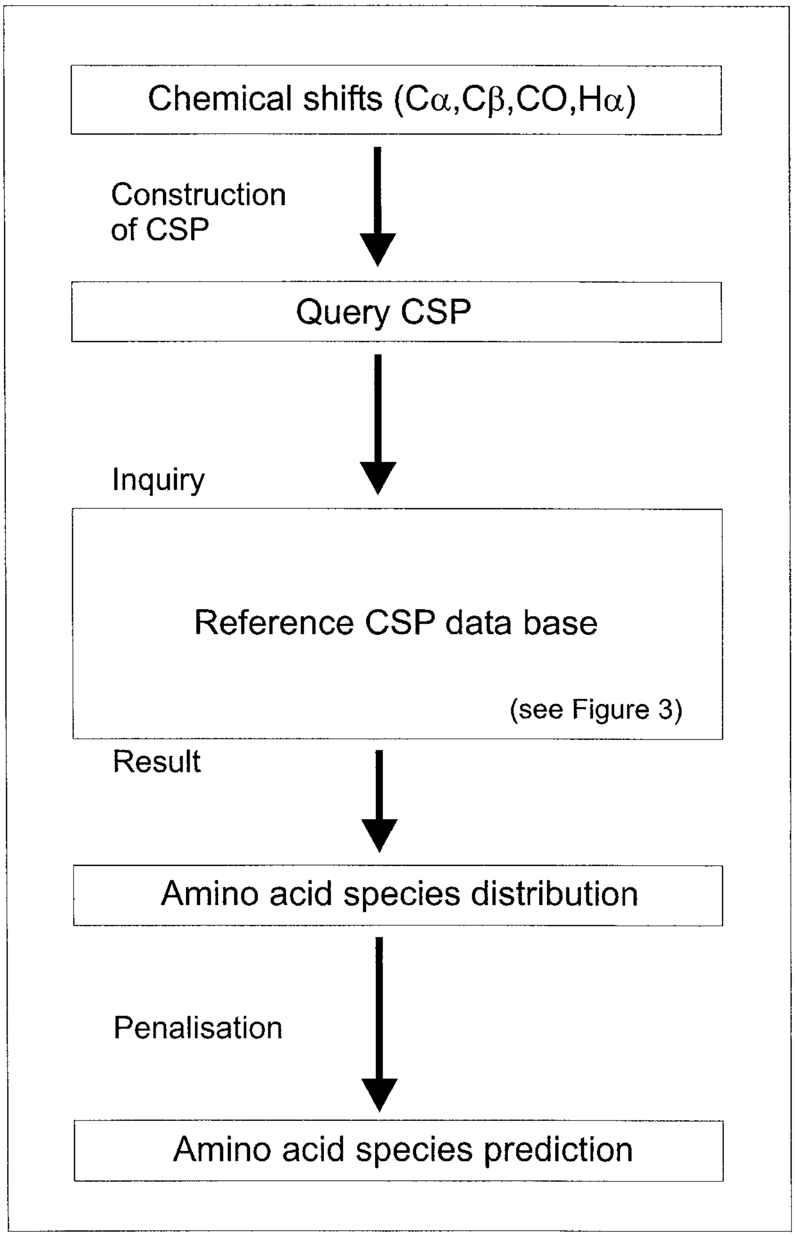

Figure 4. Flow chart of the PLATON algorithm.

liable results are in general obtained by both penalty functions.

The quality of the predictions is shown in Figure 6, by means of a histogram displaying the frequency of occurring ranking positions dependent upon on the applied penalty functions. Either 20 (penalty function 1) or 19 (penalty function 2) amino acid species were correctly predicted.

Using penalty function 1 leads to 63.5% (33 of 52) of the correct amino acid species ranked in the first three positions, whereas penalty function 2 predicts 76.9%, corresponding to 40 out of 52 amino acid species, in the first three positions. Currently, penalty function 2 yields better results than penalty function 1.

In order to verify the quality of the prediction, PLATON was tested on a further 9 proteins (Table 1) using the same parameters as in the α-spectrin SH3 do-

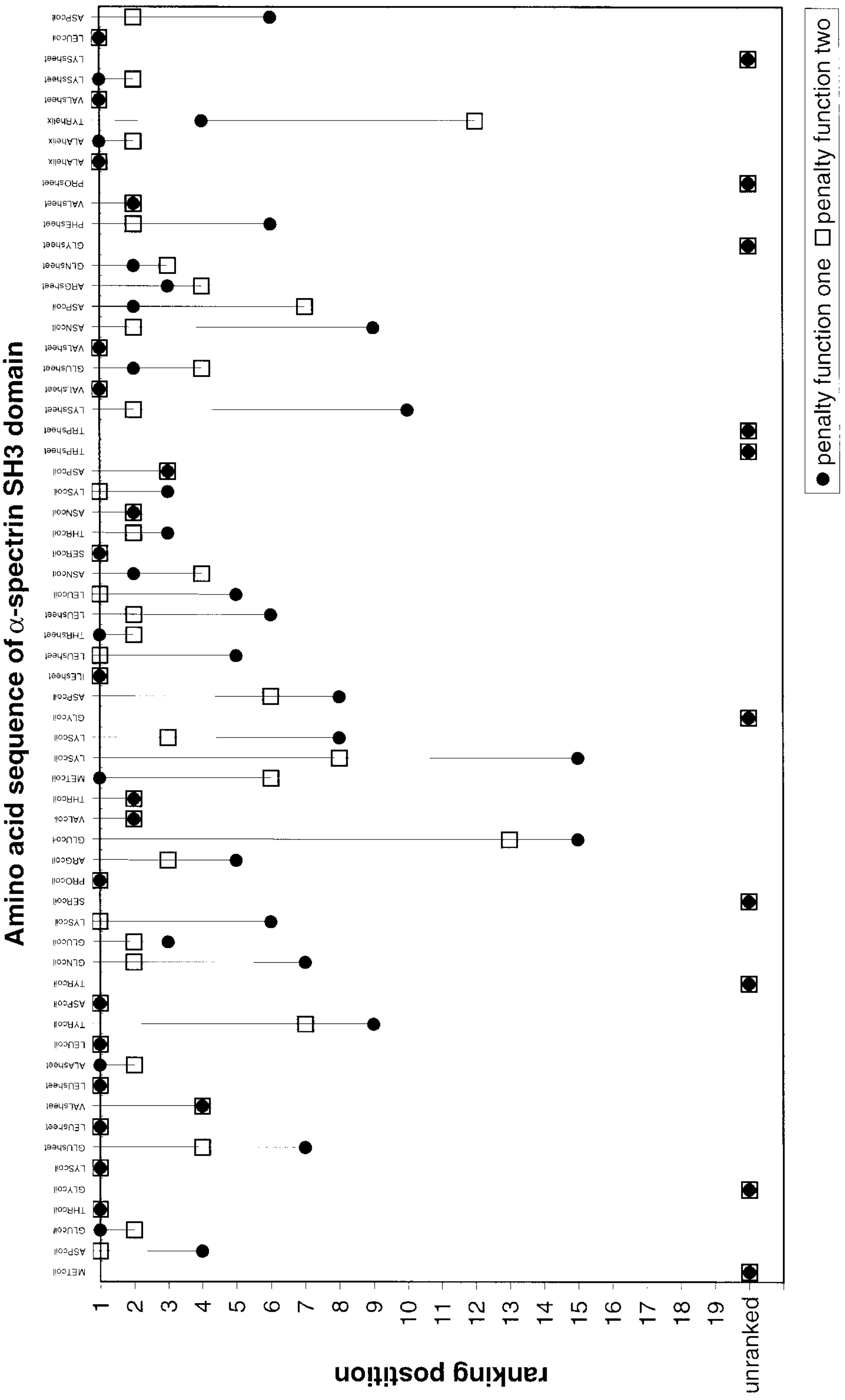

Figure 5. Ranking positions of the amino acid species of the α-spectrin SH3 domain after application of both penalty functions. An unranked position means that either the amino acid species was not contained in the distribution or the query CSP could not be created due to missing chemical shifts.

Table 1. Results of PLATON test runs on 10 selected proteins. The percentages of recognition are calculated from all amino acid species for which it was possible to create a query CSP. Depending on the applied penalty function, the first and the first three ranking positions are reported. In addition CSPs which had not the correct entries in the distribution lists are shown (number of distributions w/o correct amino acid species [%])

PDB entry	BioMagRes entry	Penalty function 1	Penalty function 2 Ranking pos. 1–3 [%] (Ranking pos. 1 [%])	PROTYP prediction (*only amino acid type*)	Number of distributions w/o correct amino acid species [%]	Secondary structure
1ev0	4237	58.7 (47.8)	73.9 (52.2)	78.3 (56.5)	13.04	2 strands 3 helices
1br0	4198	82.6 (47.0)	91.3 (74.8)	93.0 (67.0)	3.48	3 helices
1sym	4001	59.8 (43.3)	78.4 (55.7)	78.4 (49.5)	10.31	2 strands 4 helices
1dmo	4056	77.1 (46.8)	85.3 (61.5)	90.8 (56.9)	1.83	4 strands 8 helices
1qjt	4140	69.0 (40.5)	83.3 (65.5)	91.7 (64.3)	9.52	6 helices
1c0v	4146	88.1 (74.6)	93.2 (83.1)	79.7 (59.3)	0.00	2 helices
1bjx	4156	76.1 (44.3)	86.4 (56.8)	94.3 (60.2)	1.14	4 strands 4 helices
1irf	4161	63.7 (34.1)	71.4 (50.5)	82.4 (53.8)	6.59	4 strands 4 helices
1bym	4183	72.8 (30.0)	85.1 (55.0)	92.5 (66.2)	3.51	5 strands 4 helices
1aey	α-spectrin SH3 domain	63.5 (38.5)	76.9 (36.5)	78.9 (40.4)	8.77	5 strands 1 helix

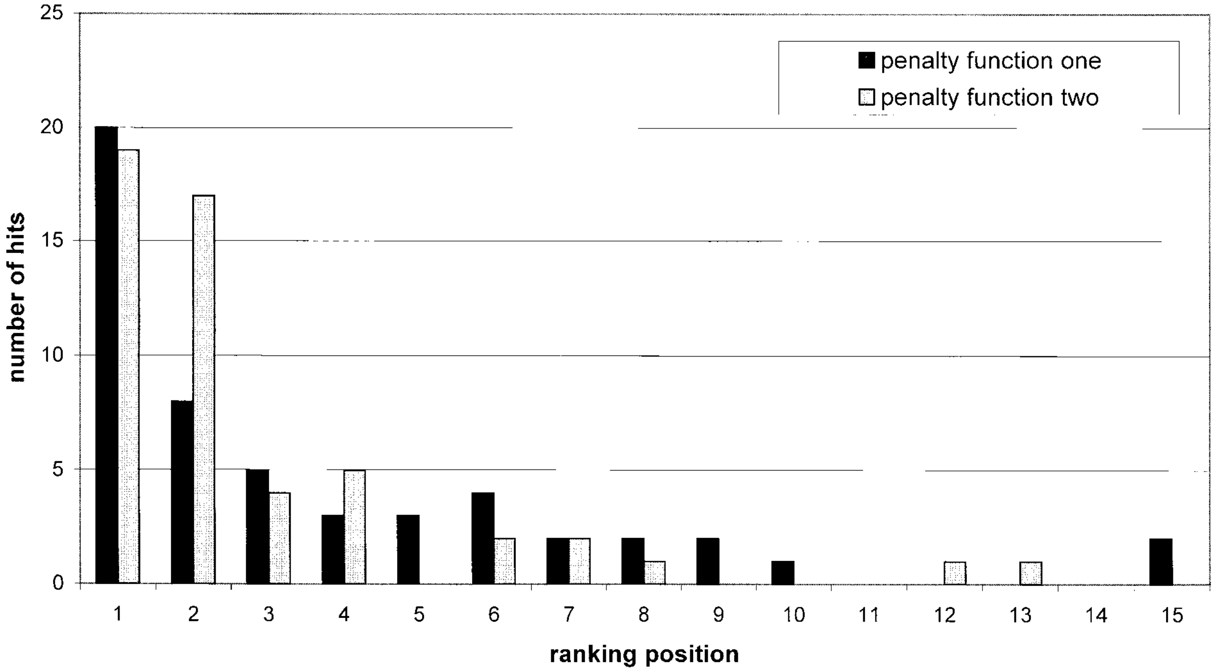

Figure 6. Calculated histogram of the ranking positions for the amino acid species of the α-spectrin SH3 domain. Upon application of penalty function 2, 76.9% of the found amino acid species are predicted in the first three positions whereas penalty function 1 predicts 63.5% in the first three positions.

main test run. The selected proteins belong to different structural classes. The results of PLATON varied depending on the chosen examples. Where only the first three ranking positions are considered, the percentage of correctly predicted amino acid species ranges from 58.7% to 88.1% for penalty function 1 and from 71.4% to 93.2% for penalty function 2. If only the first ranking position is taken as a result, the correct predictions range from 30.0% to 74.6% for penalty function 1 and from 36.5% to 83.1% for penalty function 2. An obvious trend for the ability of PLATON to predict amino acid species, dependent on their structural class, could not be observed.

We also applied PROTYP to the 10 test proteins (Table 1). For comparison PROTYP yields slightly better results if ranking positions 1–3 are taken into account. If only the first ranking positions are considered, both programs have a comparable prediction rate, although PLATON gives additional information about the secondary structure.

If PLATON is used to identify only the correct amino acid type (without secondary structure information) the performance is even further increased. This has been demonstrated for the α-spectrin SH3 domain. In the case of penalty function 1 the number of correctly predicted amino acids (ranking position one) rises from 20 to 28, penalty function 2 yields further 12 correctly predicted amino acids totaling in 31. Interestingly, 3 out of the former 5 amino acid species which were not contained in the distribution lists could be now obtained as amino acid types, namely Trp41, Trp42, and Lys60. We compared the performance of PLATON and PROTYP as modules for amino acid type prediction only. PROTYP, which is the only comparable program with similar input and output, yielded 23 correctly predicted results on ranking position one which is significantly less than the 31 top scores of PLATON in the more favorable case (data not shown). However, the performance of penalty function 1 should improve with an increasing database. If more patterns are added and, more importantly, the number of amino acid species for its reference CSP increases, the penalizing improves.

On the α-spectrin SH3 domain we tested various parameter settings for the algorithm in order to optimize its performance. Table 2 describes the percentage of correct amino acid species contained in the distributions and the average number of entries depending on the number of iterations. If one iteration is applied, all amino acid species are contained in the distributions,

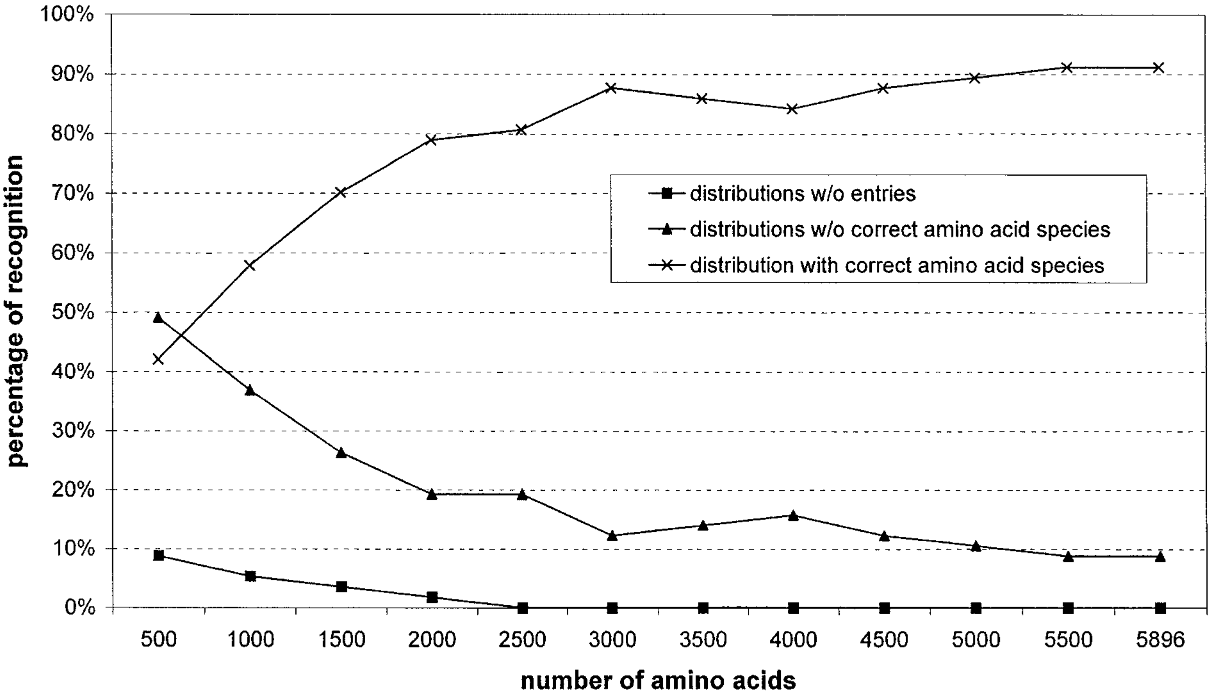

Figure 7. Illustration of the ability of PLATON to learn new reference CSPs on the example of the α-spectrin SH3 domain. The percentage of recognition is plotted vs. the amount of amino acids used as an input for the generation of the database.

Table 2. On the example of the α-spectrin SH3 domain the optimum between percentage of correct recognition and the number of possible hits are shown

Number of distributions w/o entries (%)	Number of distributions w/o correct amino acid species (%)	Number of distributions with correct amino acid species (%)	Average length of the distributions of the query CSPs	Number of iterations
0	0	100	27.2	1
0	8.8	91.2	6.1	2
17.5	38.6	43.9	2.2	3

however with high ambiguity. The average number of entries in the distributions is 27.2 per query CSPs. In the case of three iterations, both values decrease. The distributions are now less ambiguous (2.2 entries per query CSP) but either the number of distributions without entries increases from 0 to 17.5%, or the number of distributions not containing the correct amino acid species increases from 0 to 38.6%. Thus, two iterations are currently applied as the optimal compromise for the existing amount of chemical shifts in the reference CSP database and performance of PLATON. For the 57 analyzed residues of the α-spectrin SH3 domain 91.2% of the amino acid species were correctly contained in the distributions and the average length of the distributions is 6.1.

Apart from increasing iteration levels, the result could be further improved by enlarging the database. The reliability of recognition, that is the correct amino acid species is contained in the distribution, increased with the number of amino acids contained in the reference CSP database. In Figure 7, the percentage of recognition is plotted against the number of amino acids used as input for the generation of the reference CSP database. With currently 5896 chemical shift sets in the database, and without having added the α-spectrin SH3 domain chemical shifts, above 90% of the α-spectrin SH3 domain amino acid species are now elements in the distributions.

If the reference CSP database is constructed either from CO, Cα, and Cβ, or from Cα, Cβ, and Hα chemical shift values, the percentage of recognition further increases up to 98.25% (data not shown) but the average length of the distributions elongates from 6.1 to 12.6. The input of a higher number of chemical shifts improves the reliability of the predicted amino acid species significantly.

After resonance assignments and the structural work of the investigated proteins are complete, newly created CSPs can be entered into the database, which is hence constantly growing. In this way, query CSPs which were not associated with an amino acid species before are added to the reference CSP database increasing the reliability of PLATON. If the number of successful prediction hits increases with increasing the database, it would be feasible to introduce another iteration step. However, given the current number of database entries, only two iteration steps are recommended.

The reference CSP database and the source code of the program PLATON are available under http://www.fmp-berlin.de/∼labudde. PLATON was written in C and the graphical user interface was created with Tcl/Tk.

Conclusion

In this work, we have used PLATON to predict amino acid species from unassigned chemical shift values, which was demonstrated for the α-spectrin SH3 domain and tested on a further 9 proteins (Table 1). Our results show that it is possible to use CSPs for the prediction of amino acid species. The percentage of the correct amino acid species occurring in the first three positions in the ranking list, ranges from 71.4% to 93.2% using the more favorable penalty function. If only the top result of the ranking list for these 10 proteins is considered, 36.5% to 83.1% of the amino acid species are correctly predicted.

The main advantage of our approach over other methods which rely on for example average chemical shift values, is the ability of the reference CSP database to extend the distributions for reference CSPs, and thus to improve significantly its performance with increasing content. This has two consequences for the quality of the prediction: First, untypical query CSPs can be learnt after the verification and the reliability of the prediction of underrepresented reference CSPs is improved. Second, as outlined in Table 2, the amount of data limits the number of iterations used to construct the reference CSPs, implying that more data would allow the introduction of further subdivision steps and therefore to build more specific reference CSPs.

Furthermore, we could demonstrate that unassigned chemical shifts derived from standard triple resonance experiments are a suitable starting point for predicting amino acid species. Data input consists solely of backbone resonances and Cβ chemical shifts; no use of homonuclear side chain chemical shift information was made. PLATON yields results for each amino acid type, without the need for forming amino acid classes.

Also, the number of considered secondary structure elements could be increased, for example by splitting the secondary structure class coil in turns, loops, and hairpins or by differentiating between different helix types. In order to improve the reliability of the secondary structure prediction, the secondary structure of sequential neighbors could be taken into account for choosing the correct entry on top of the ranking list e.g. by considering three residues. The use of higher dimensional chemical shift spaces is also conceivable and should lead to more reliable predictions.

We are currently working on including PLATON into a complete sequence mapping program with simultaneous secondary structure element prediction. At present we use PLATON in combination with the resonance assignment program CATCH23 (Croft et al., 1997) to evaluate pattern search results. It is also feasible to use PLATON to verify assignments which are produced in the course of a classical manual procedure.

Acknowledgements

We thank A. Steuer for secretarial assistance. We are grateful for the financial support of this work by the BMBF-Leitprojekt ‘Strukturanalyse mit hohem Durchsatz für medizinisch relevante Proteine – Proteinstrukturfabrik’ (Fk. 01GG9812).

References

Bartels, C., Güntert, P., Billeter, M. and Wüthrich, K. (1997) *J. Comput. Chem.*, **18**, 139–149.

Bernstein, F.C., Koetzle, T.F., Williams, G.J.B., Meyer, E.F., Brice, M.D., Rodgers, J.R., Kennard, O. Shhimanouchi, T. and Tasumi, M. (1977) *J. Mol. Biol.*, **112**, 535–542.

Blanco, F.J., Ortiz, A.R. and Serrano, L. (1997) *J. Biomol. NMR*, **9**, 347–357.

Chou, P.Y. and Fasman, G.D. (1974) *Biochemistry*, **13**, 222–245.

Cornilescu, G., Delaglio, F. and Bax, A. (1999) *J. Biomol. NMR*, **13**, 289–302.

Croft, D., Kemmink, J., Neidig, K.P. and Oschkinat H. (1997) *J. Biomol. NMR*, **10**, 207–219.

Fasman, G.D. (1989) *Trends Biochem. Sci.*, **14**, 101–162.

Grzesiek, S. and Bax, A. (1993) *J. Biomol. NMR*, **3**, 185–204.

Heinemann, U., Frevert, J., Hofmann, K., Illing, G., Maurer, C., Oschkinat, H. and Saenger, W. (2000) *Prog. Biophys. Mol. Biol.*, **73**, 347–362.

Huang, K., Andrec, M., Heald, S., Blake, P. and Prestegard, J.H. (1997) *J. Biomol. NMR*, **10**, 45–52.

Kay, L.E. (1997) *Biochem. Cell Biol.*, **75**, 1–15.

Kyngäs, J. and Valjakka, J. (1998) *Protein Engin.*, **11**, 345–348.

Leutner, M., Gschwind, R.M., Liermann, J., Schwartz, C., Gemmecker, G. and Kessler, H. (1998) *J. Biomol. NMR*, **11**, 31–43.

Medek, A., Olejniczak, E.T., Meadows, R.P. and Fesik, S.W. (2000) *J. Biomol. NMR*, **18**, 229–238.

Moseley, H.N.B. and Montelione, G.T. (1999) *Curr. Opin. Struct. Biol.*, **9**, 635–642.

Moseley, H.N.B., Monleon, D. and Montelione, G.T. (2001) *Meth. Enzymol.*, **339**, 91–108.

Murzin, A.G., Brenner S.E., Hubbard T., and Chothia C. (1995) *J. Mol. Biol.*, **247**, 536–540.

Musacchio, A., Saraste, M. and Wilmanns, M. (1994) *Nat. Struct. Biol.*, **1**, 489–491.

Pons, J.L. and Delsuc, M.A. (1999) *J. Biomol. NMR*, **15**, 15–26.

Seavey, B.R., Farr, E.A., Westler, W.M. and Markley, J. (1991) *J. Biomol. NMR*, **1**, 217–236.

Wishart, D.S. and Case, D.A. (2001) *Meth. Enzymol.*, **338**, 3–34.

Wishart, D.S. and Nip, A.M. (1998) *Biochem. Cell Biol.*, **76**, 153–163.

Wishart, D.S. and Sykes, B.D. (1994) *J. Biomol. NMR*, **4**, 171–180.