CRAACK: Consensus Program for NMR Amino Acid Type Assignment

Cindy Benod, Marc-André Delsuc, and Jean-Luc Pons*

Centre de Biochimie Structurale, CNRS UMR 5048, INSERM UMR 554, Université Montpellier 1, 29 rue de Navacelles, 34090 Montpellier, France

Received March 22, 2005

Protein peak spectrum assignment is a prerequisite of the nuclear magnetic resonance study of a molecule. We present here a computer tool which proposes the determination of the amino acid type from the values of the chemical shifts. This tool is based on two consensus algorithms based on several published typing algorithms and was trained and extensively tested against the Biological Magnetic Resonance Bank chemical shift data bank. The first one accomplishes the analysis with support vector machine technology, grouping related amino acids together, and presents a mean rate of success above 90% on the test set. The second one uses a classical consensus algorithm of vote. Furthermore, secondary structural prediction is available. This tool can be used for assisting manual assignment of peptides and proteins and can also be used as a step in an automated approach to assignment. This program has been called CRAACK and is publicly available at the following URL: http://abcis.cbs.cnrs.fr/craack.

INTRODUCTION

Nuclear magnetic resonance (NMR) is a very powerful tool for the study of proteins in solution, and the structure determination of biomolecules starts generally with spectral assignment. The first step is the determination of resonance frequencies (or chemical shifts) of the signals observed in the different acquired spectra. The assignment of the observed lines permits the interpretation of the observed phenomena at the atomic level. Furthermore, assignment is a requisite for structural determination. Unfortunately, despite the recent progresses obtained thanks to the use of isotopic labeling of the protein backbone, the assignment task[1] is still a difficult one.

To ease this assignment procedure, several groups have developed methods in order to automate the assignment.[2−4] In addition to these approaches permitting the proposal of the complete assignment from the complete analysis of the set NMR spectra, several authors have proposed a partial approach consisting of the assignment of the amino acid type from the measured chemical shifts only. This operation, which we will call typing, only attempts to assign the amino acid type, with no reference to any sequential assignment. Several approaches have been proposed in this direction, from the early works of Protyp[5] or Rescue[6] to more-recent developments such as Platon,[7] RescueN,[8] and Rescue2.[9] Additionally, a previously unpublished typing program, developed in our group and nicknamed SVMTyping, is presented.

The typing approach has been made possible thanks to the development of large chemical shift databanks and, more specifically, the Biological Magnetic Resonance Bank (BMRB).[10] The more-recent developments usually profit from the extension of the available databanks and, consequently, present results of higher quality. However, the mathematical foundations are usually different, and it would be interesting to compare the results generated by the different programs.

This work presents results in this direction. We first compare the typing programs available on the basis of their mathematical foundation, required information, and overall quality of the prediction process. It appears that none of the tested programs present a quality level sufficient to permit the unambiguous assignment of the amino acid types. On the other hand, these different programs rely on quite different mathematical modelizations of the problem and require different sets of measured chemical shifts. A detailed analysis of the outcome of each method shows that part of the assignment errors are systematically found in all of the programs, while some errors are specifically found in a single one.

A consensus approach of the typing problem is presented here. The aim is to optimize the prediction step by combining the output of existing typing modules while detecting errors specific to one or a few modules. As an additional benefit, the quality of the consensus analysis can be estimated from the discrepancy between the different typing modules. This consensus approach is possible today because of the size of the BMRB. The number of chemical shifts deposited permits the building of large training sets for the learning algorithms to be implemented while, at the same time, using large test sets for the validation of these algorithms.

It should be noted that the consensus approach, widely used for protein primary sequence analysis as well as secondary and tertiary structure prediction, has barely been used for the NMR assignment problem.

To build the consensus tools, we chose to rely on a kind of algorithm that models the problem on a set of nonlinear kernel functions. Among the kernel methods, the support vector machines (SVM) technique[11−14] was chosen. In its current implementation, SVM builds a nonlinear modelization of the analytical problem on the basis of a set of nonlinear kernel functions. SVM is a powerful methodology for solving problems in nonlinear classification, function estimation, and density estimation, which has also led to many other recent developments in kernel-based methods in general. Originally, it was introduced within the context

* Corresponding author e-mail: JL.Pons@cbs.cnrs.fr.

10.1021/ci050092h CCC: $33.50
Published on Web 02/21/2006

of statistical learning theory and structural risk minimization. In this approach, one has to solve convex optimization problems, typically by quadratic programming. The minimization is usually based on the insensitive loss function, which permits some margin around the searched optimum.

The SVM approach permits easily handling of large input vectors, as well as numerous missing values. Additionally, SVM affords a very fast learning curve even on large training databases. In these respects, SVM is superior to the neural network approach to which it can be compared.

SVM was found to be efficient for the consensus typing problem with relative amino acids grouped together. A second approach to the consensus problem is also presented here, based on a more classical vote strategy. This strategy proved to be efficient for a finer analysis on a single amino acid basis. Both approaches were combined in order to build a complete typing tool. This tool has been nicknamed CRAACK, which stands for consensus rules for amino acid characterization using a kernel method, and is freely available as an online program at the following URL: http://abcis.cbs.cnrs.fr/craack.

MATERIALS AND METHODS

Creation of Test and Training Bases. The learning and consensus algorithms used in this work were trained and tested on a set of chemical shifts. The test and training bases have been created from the BMRB[10] (downloaded in January 2004). Spins systems have been extracted from STAR files according to the two criteria that proteins or peptides should have more than 10 residues and that the greater part of ^{1}H, ^{13}C, and ^{15}N chemical shifts should be assigned (determined from the ratio of chemical shifts to the number of residues greater than five).

Chemical shifts were compared to the limit values as defined by Marin et al.[9] in a statistical analysis on 783 nonhomologous sequences of the BMRB. Chemical shifts which exceed these limits were not taken into account. For the same reasons, proteins containing a paramagnetic center were not taken into account.

It was verified that consideration of pH value, temperature value, or NMR reference had no effect on the quality of the final results, and as a consequence, this information was not used in this work.

Two independent sets of chemical shifts were constructed from the BMRB: the first one is based on all of the amino acids presenting at least chemical shift values for the HN, N, and HA spins; the second represents the prolines and the amino acids for which the HN value is missing.

The first set, devoted to nonprolines, was built in the following manner. Out of the BMRB entries, selected as described above, amino acids were randomly chosen and the observed chemical shift values entered in the database. The same number of spins systems was extracted for every residue kind.

Finally, chemical shifts corresponding to 3800 spin systems (200 spin systems per residue) were extracted from the BMRB. The selected spin systems were randomly separated into two equivalent parts.

The first part is used to create a training base (called Lbase: 60 spin systems per residue) and a test base (called Tbase: 40 spin systems per residue). These two bases are used for the optimization of the two consensus tools (SVM strategy and vote strategy). To create a base with incomplete spin systems, the same spin system is represented four times with the following chemical shifts: HN/N/HA/HB/CA/CB, HN/N/HA/HB/CA, HN/N/HA/HB, and HN/N/HA/CA/CB. The reason for entries with missing values is to simulate the presence of incomplete spectra, to increase the robustness of the tool when analyzing partial spectra.

The second part of the selected spin systems is used to create a second test base. This base, called TestDBc (1900 complete spin system, 100 spin systems per residue), is used to test all of the tools which are analyzed in this work and to validate the choices made during the optimization. We chose to use two test bases for the optimization of training and for validation in order to minimize the possibility of overfitting of the data.[6] As with the other bases (shown before), TestDBc had been filled in with incomplete spin systems in order to create TestDB, which has 7600 spin systems (four levels of incompleteness for 100 spin systems per residue).

The second set, devoted to prolines, is built in the following manner. A total of 700 amino acids not presenting the HN chemical shift were extracted from the BMRB, representing 400 prolines and 300 amino acids of another kind. In a way similar to above, this set was separated in several bases: LBaseP contains 300 prolines and 300 nonproline residues and is devoted to the training of the different tools; the other 100 prolines were injected into TestDBc and TestDB in order to validate CRAACK. Finally, TestDBc contains 2000 entries (100 spin systems per residue), and TestDB contains 8000 entries (with four levels of incompleteness per spin system).

Furthermore, all of the typing tools, as well as the consensus strategies, were evaluated against the RefDB database.[15] This database is constructed by filtering out problematic entries in the BMRB and by correcting all of the chemical shift values depending on the chemical shift reference and pH used during the study. For robustness reasons, we trained the tools developed here on uncorrected chemical shift values extracted from the BMRB. However, we also chose to evaluate these tools on the RefDB, using this database as a golden standard of chemical shift variations.

Existing Typing Modules. Four typing tools described in the literature, and one additional tool which was not previously published, were used for the consensus strategy.

All of the typing tools presented are based on the same principle: starting from a more or less complete list of chemical shifts belonging to the same residue, a target amino acid is proposed, eventually with an associated confidence level. These tools can be separated in two classes: tools in the first class propose one given amino acid, or a list of amino acids, with decreasing levels of confidence, whereas tools in the second one put similar amino acids together into groups and only determine which group the amino acid belongs to.

The different typing tools used for this work are presented in Table 1. The quality of the prediction shown is estimated on the basis of the percentage of correct answers, for three test bases: TestDBc (complete), TestDB (incomplete in terms of chemical shifts), and Ref DB.

Table 1. Rate of Success of Classic Typing Module[a]

typing tools	strategy	selective capacity	rate of success on TestDBc (2000 tests)	rate of success on TestDB (8000 tests)	rate of success on RefDB (19 299 tests)
Rescue	neural network	10 groups	57.50% (1986 answers)	58.11% (6054 answers)	50.52% (16846 answers)
RescueN	neural network	7 groups	88.37% (1900 answers)	73.07% (7600 answers)	78.09% (19003 answers)
Rescue2	Bayesian algorithm	20 residues	71.47% (1998 answers)	54.30% (7782 answers)	71.61% (19247 answers)
Platon	discriminant analysis	20 residues	51.99% (1662 answers)	51.48% (3265 answers)	58.17% (14325 answers)
SVMTyping	support vector machine	19 residues	68.57% (1893 answers)	56.87% (6691 answers)	68.61% (18394 answers)

[a]Sometimes the module does not give an answer. The number of valid answers is given in parenthesis.

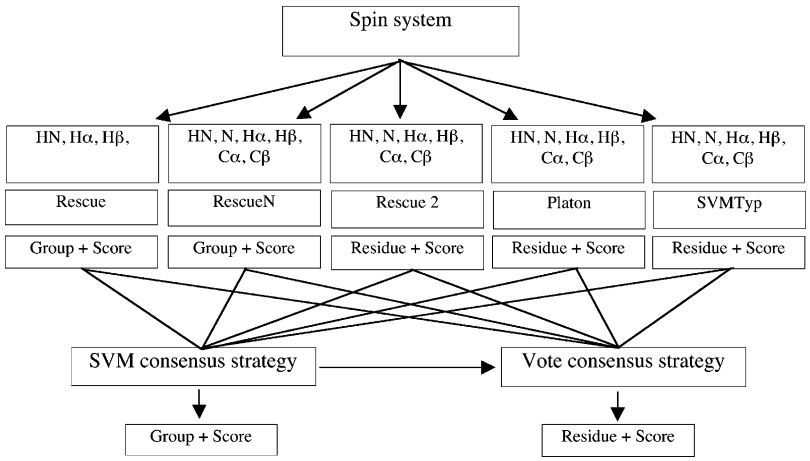

Figure 1. Flowchart of CRAACK. A studied spin system is input into the five typing modules, which generate five different answers. These answers are used by two consensus strategies (SVM and classic vote algorithm) in order to propose a global typing solution.

It can be observed that approaches which regroup similar amino acids attain higher results. The lower values observed for Rescue[6] are explained by the fact that this method only considers proton chemical shifts.

These typing tools are based on different mathematical approaches. Neural network algorithms are used in Rescue[6] and RescueN.[8] Rescue[6] uses proton chemical shifts as input (originally, Rescue used all available proton chemical shifts, but only HN, HA, and HB were considered here) and proposes a classification in 10 groups of amino acids. It offers 57.50% correct answers. RescueN[8] is based on the same principle and uses ^{1}H and ^{15}N chemical shifts as input. The rate of success is higher (88.37%) to the detriment of a weaker dividing power (separation in seven groups only).

The typing method Rescue2[9] uses a Bayesian statistical analysis technique. It uses all of the available chemical shifts (^{1}H, ^{15}N, and ^{13}C) and offers a complete separation of the 20 residues and a score of 71.47%.

Platon software[7] uses a discriminative analysis strategy. It allows the separation of 20 residue types with the help of ^{1}H, ^{15}N, and ^{13}C chemical shifts. In our tests, the score (51.99%) seems to suffer from the absence of CO chemical shifts, unavailable in our test base. Furthermore, Platon proposes secondary structure predictions.

The tool Protyp[5] was not found in exploitable shape and was not used in this work.

A previously unpublished piece of typing software, previously developed in our laboratory, was also used in this study. This tool, called SVMTyping, is based on the SVM approach.[11] It allows the typing of 19 amino acids residues from ^{1}H, ^{15}N, and ^{13}C chemical shifts and does not handle entries lacking values for HN spins (thus, it does not handle proline). To increase the robustness of this tool, a training base (with 1050 spin systems) was built with seven levels of incompleteness (HN/N/HA/HB/CA/CB/CO, HN/N/HA/HB/CA/CB, HN/N/HA/HB/CA, HN/N/HA/HB, HN/N/HA/CA/CB/CO, HN/N/HA/CA/CB, and HN/N/HA/CA). Spin systems for which the chemical shift value is missing for the HN, N, or HA spins are not considered. The test base contained 1900 spin systems (100 per residue type).

SVMTyping was built using the SVMTorch II software,[16] which is an implementation of SVM. SVMTorch was used in sparse mode, with a set of Gaussian kernel functions (-t option set to 2) and a width of the error pipe (-eps option) set to 0.005. A total of 19 different models (one for every amino acid except proline) were built.

From the given input, the output of each of the 19 models is sorted with decreasing levels of preference. Additionally, a reliability is given for every result by computing the distance from the location of the residue point to the dividing hyperplane, calculated in the hyperspace. After optimization, this tool offers 64.2% correct answers on the SVMTyping training base and 66.84% on the SVMTyping test base (82.58% in the first three ranks).

For comparison with other typing modules, the SVMTyping rate of correct answers on the TestDBc base is 68.57%.

Technical Organization of CRAACK. Figure 1 presents the general flowchart of the CRAACK method, as presented in this work. All of the chemical shifts available for a given spin system are directed, with the proper format, to each typing module used for the consensus approach. Each module generates an answer, and this set of answers is used as input by the consensus algorithms.

First Approach: Consensus by SVM. The first strategy is based on support vector machines.[11] The MySVM toolbox software[17] was chosen for this purpose because it allows coding of the entries with a high number of missing values.

Table 2. Rate of Success of the Consensus Typing Modules

typing tools	strategy	selective capacity	rate of success on TestDBc (2000 tests)	rate of success on TestDB (8000 tests)	rate of success on RefDB (19299 tests)
consensus by SVM	support vector machine	8 groups	95.43% (1924 answers)	87.77% (7610 answers)	91.7% (18 206 answers)
consensus by vote	vote algorithm	20 residues	73% (2000 answers)	54.53% (8000 answers)	71.09% (19 299 answers)

It is essential to design an optimum coding of the typing modules output, as an input vector for the SVM consensus step. This input vector should contain the information from the five modules, including the predicted residues and their associated scores. This coding has to bring all of the necessary information for the final residue classification without bringing unwanted noise. Several types of input coding were tested. The most-effective coding consists of 84 elements (corresponding to the total number of possible answers for the five modules: 18 for Rescue with a combination of its two neural networks,[6] 7 for RescueN, 20 for Rescue2 and Platon, and 19 for SVMTyping). The input vector for the SVM analysis contains the score values associated with the answers of each module. The score values are restricted to the three residues with the highest scores for each module, and the remaining answers are set as missing values. These scores were normalized in order to permit a distinction between the first, second, and third ranks.

Among the set of tested SVM kernel functions, the dot function obtained the best predicting scores.

A different SVM model was optimized for every possible output of the SVM analysis. Three levels of amino acid grouping were tested (full separation in 20 residues, partial separation in 11 groups, and partial separation in eight groups). The groups were chosen according to similarities in the residue profiles. Separation in 20 residues only gives a weak score of 39.80% correct answers on the TestDB base. On the same base, separation in 11 groups (A, T, V, G, L, I, H, FYWCH, KMEQR, DN, and P) gives 72.11% correct answers, and separation in eight groups (A, H, T, G, VI, FYHCWDN, and EQMKRL P) gives 87.77% correct answers (Table 2).

Second Approach: Consensus by Vote. Another approach evaluated here consists of using the vote technique generally used in solving consensus problems. Only the three answers with the highest scores for every module are considered for the vote. Every residue predicted by the five modules is given a vote. The votes are weighted with optimized values, which depend on the typing module (overall reliability of the module) and the residue answer rank. The weight values are shown in Table 3. For every residue, weighted votes are added. The amino acid residue obtaining the highest vote is chosen.

The rate of success is increased when the result of the consensus tool based on the SVM (eight groups) is added in this voting process. The rate of correct answers for the TestDB basis is 54.53% (Table 1). The consensus by vote is able to separate the 20 residue types. The number of votes obtained for each residue is used as a reliability descriptor.

Prediction of Secondary Structures. Once the amino acid residue is determined, the secondary structure of this amino acid may be predicted from the values of chemical shifts, using the chemical shift index (CSI) approach.[18−19] Results of the CSI analysis are presented here, using HA, CA, CB, and CO chemical shifts and residue type information.

Table 3. Weight of Typing Module Answers after Optimization

typing tools	weight of answer
Rescue	1.2 for score larger than or equal to 90.
(for all residues	1 for score between 0 and 90.
of chosen group)	0.8 for score equal to 0.
RescueN	1.2 for score larger than or equal to 60.
(for all residues	1 for score between 35 and 60.
of chosen group)	0.8 for score smaller than or equal to 35.
	1.2 for score larger than or equal to 60.
Rescue2 (first choice)	1 for score between 35 and 60.
	0.8 for score smaller than or equal to 35.
Rescue2 (second choice)	0.5 for all scores
Rescue2 (third choice)	0.25 for all scores
Platon (first choice)	0.8 for all scores
Platon (second choice)	0.5 for all scores
Platon (third choice)	0.25 for all scores
SVMTyping	1.2 for all scores
	1.2 for score larger than or equal to 1.
SVM-consensus	1 for score between 0.5 and 1.
	0.8 for score smaller than or equal to 0.5.

RESULTS AND DISCUSSION

The results obtained with the two consensus methods are shown in Table 2 and are detailed in Figure 2.

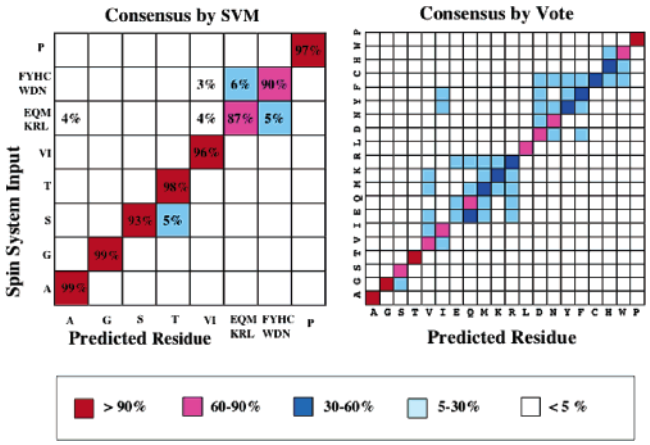

Figure 2. Response of the two consensus strategies as observed on the RefDB test base. For each family of amino acid analyzed (located on the left), the number of answers is given graphically (by color). The rate of success for the vote is 71.09%. The rate of success for the SVM consensus is 91.7%. In the case of the SVM consensus, the scores are written in the grid.

Table 2 presents the global efficiency of the two consensus methods. It can be seen that the SVM tool, separating the residue types into eight classes, is more efficient than any other single typing tool with the same selective capacity. Similarly, the vote tool, with a classification of all of the amino acids, is also more efficient than any other tool. Figure 2 details the residual errors obtained with both consensus methods.

The different modules used for typing prediction have different rates of failure and fail to give the correct answer for different reasons, and in different ways. Even though most of them are optimized on training sets extracted from the same database (the BMRB), they are implemented using very

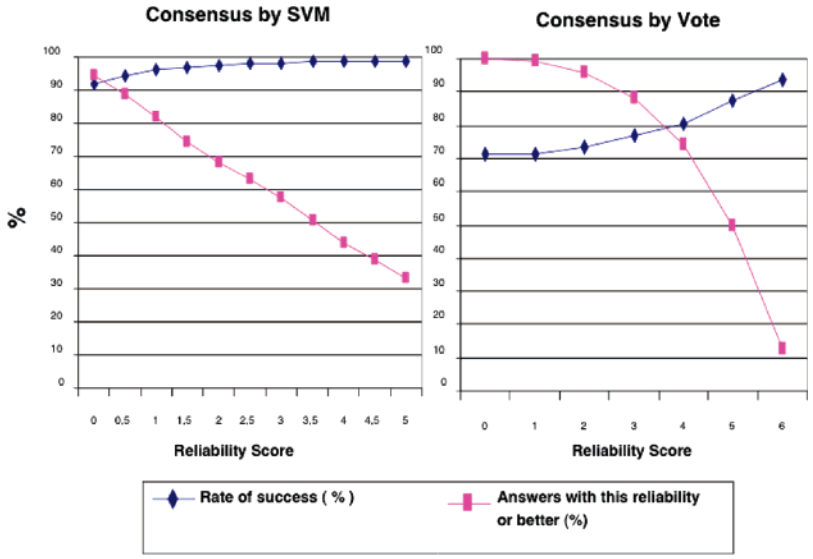

Figure 3. Rate of success on TestDBc for all the answers over a given reliability coefficient (abscissa). The percentage of entries for a given reliability is also given. In the case of the SVM strategy, only answers with positive reliability are taken into account.

different mathematical strategies, which react differently. It is for this reason that the consensus approach is pertinent.

Indeed, a detailed analysis shows that many systematic errors, specific to one approach, are corrected by the consensus tools. On the other hand, systematic errors, common to all of the prediction modules, remain after the consensus step.

A detailed analysis brings the following points: The separations between F and Y and between M and K are fuzzy, and these residues are quite systematically mistyped. Residues E and D are systematically underpredicted, in favor of Q and N. This is probably due to the large similarity of the side chain presented by these residues. Additionally, C is systematically underpredicted in favor of W. This may be due to the "atypical" pattern presented by C, which can be found in two different chemical states: oxidized or reduced. Chemical shift variations with respect to the oxidation state of C have already been described.[20] The variation of the oxidation state, which is not coded in this study (and not in the cited studies either), may thus blur the distinction between C and the other AMX spin systems.

The different typing modules are not sensitive to the same missing chemical shift values in the spin systems. HB is important for the performance of Rescue and RescueN. On the other hand, Rescue2 and SVMTyping are more sensitive to the lack of CB, and Platon is sensitive to the lack of CO. These differences in sensitivity are obviously an opportunity for a consensus approach, and one may hope that the consensus will be more robust in regard to missing values than any single method.

To verify that each module brings valuable information to the problem, the same calculation was performed by alternately removing each of the modules. We could observe that whichever module is removed from the computation, the consensus prediction loses some of its efficiency. Thus, every module globally brings more information than noise, and all of the modules are important for the quality of the output.

A reliability descriptor is computed for each output of CRAACK. In the case of the approach by SVM, this reliability description is the distance from the point calculated in the hyperspace to the dividing hyperplane.[11] The larger the distance, the more reliable the answer. Figure 3 shows the rate of correct answers observed for different levels of reliability. In the SVM approach, the rate of success is 91.7 if only the answers with positive reliability are taken into account. A total of 68.4% of the test base is predicted with a reliability superior to 2. The rate of success in this part of the base is then 97.5%. Despite a limited dividing power (eight groups), this tool has excellent performance.

Table 4. Rate of Success of Typing Consensus and Secondary Structure Prediction for Five Proteins Not Present in the Test and Learning Bases

protein name (PDBid)[21]	secondary structure	% SVM	% vote	% secondary structure[a]
human kinase B protein (1P6S)[22]	sheet + helix	75.76	67.96	67.14
MPS one binder (1R3B)[23]	sheet + helix	77.11	68.67	66.15
focal adhesion kinase (1QVX)[24]	helix	91.59	77.57	57.95
Cd44 antigen (1POZ)[25]	sheet + helix	77.16	69.29	84.69
superoxide dismutase (1RK7)[26]	sheet	81.64	67.09	43.59

[a] Random coil zones were not taken into account in the percentage of correct secondary structure prediction.

In the case of the vote consensus, reliability was computed from the number of votes in the process. Values go from 0 to 6.8. Figure 3 shows that more than 75% of the test base is predicted with a reliability superior or equal to 4. The rate of correct answers in this part of the base is more than 80%.

Five supplementary proteins were chosen to test the CRAACK procedure. These proteins are not present in the test bases used for the optimization of the consensus. They are described in Table 4 and are composed of α helices and β sheets in equivalent proportions, and all of them are ^{15}N- and ^{13}C-labeled. The rate of correct answers in residue typing is between 75.76% and 91.59% for the consensus by SVM and between 67.09% and 77.57% for the consensus by vote.

The secondary structure rates of correct prediction are indicated in Table 4. Coil zones were not taken into account for the computation of the rates. The results for the secondary

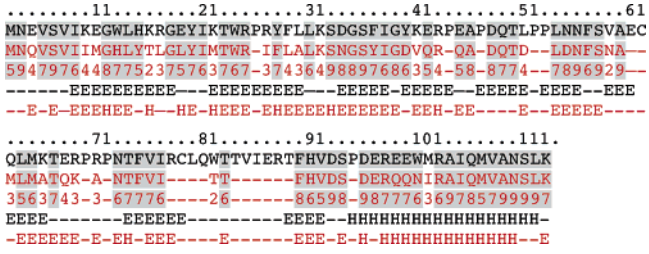

Figure 4. Typing prediction (vote strategy) and secondary structure prediction for human kinase B protein. The actual primary sequence and secondary structure are in black. Predictions are shown in red. The secondary structure is defined with the PSEA software.[27] The consensus vote reliability (0−9) is shown (below typing prediction), and the gray zones indicate prediction with a reliability larger or equal to 5.

structure prediction are fair albeit based on the approximate results of the typing prediction step. It can be observed in the five tested proteins that the quality of the secondary structure prediction is variable and does not seem to be related to the state, helix or sheet, of the protein.

Figure 4 more exactly shows the case of the human kinase B protein. The typing step and the secondary structure prediction are shown. It is noteworthy that errors are very often associated with a low typing reliability. On the contrary, for reliabilities larger than 5 (on a scale from 0 to 9), only 13 errors can be observed, among which eight are issues that have already been described between Y and F, D and N, and E and Q.

CONCLUSION

The traditional approach to the assignment of protein NMR spectra consists of assigning amino acid signals from sequential information obtained either from nuclear Overhauser effect spectrometry homonuclear experiments or from triple-resonance experiments. On the other hand, amino acid typing is a simple step which tries to determine the amino acid type regardless of the sequential assignment.

In this work, we have explored the capacities of several published typing programs and have shown that a consensus strategy is possible. We propose a consensus tool, called CRAACK, freely available on Internet, and show that this tool permits reliable amino acid typing.

The quality of the typing is dependent on the size and quality of the database on which the training is performed. The BMRB is a key component of this study. With the unavoidable improvement of the BMRB over the coming years, it can be ascertained that typing tools, such as the one presented here, will still improve their quality and prediction power. With the quality of amino acid typing already attained today, it can be anticipated that new NMR analysis procedures, relying more heavily on chemical shift information, will appear.

ACKNOWLEDGMENT

The authors acknowledge ACI-IMPBio of CNRS, INSERM, and UM1 for financial support.

REFERENCES AND NOTES

(1) Jakobsen, K.; Sletner, S.; Aalen, R. B.; Bosnes, M.; Alexander, D.; Wüthrich, K. *NMR of Proteins and Nucleic Acids*; Wiley: New York, 1986; pp 117−161.

(2) Moseley, H. N. B.; Montelione, G. T. Automated Analysis of NMR Assignments and Structures for Proteins. *Curr. Opin. Struct. Biol.* **1999**, *9*, 635−642.

(3) Altieri, A. S.; Byrd, R. A. Automation of NMR Structure Determination of Proteins. *Curr. Opin. Struct. Biol.* **2004**, *14*, 547−553.

(4) Baran, M. C.; Huang, Y. J.; Moseley H. N. B.; Montelione G. T. Automated Analysis of Protein NMR Assignment and Structures. *Chem. Rev.* **2004**, *104*, 3541−3555.

(5) Grzesiek, S.; Bax, A. Amino Acid Type Determination in the Sequential Assignment Procedure of Uniformly 13C/15N-Enriched Proteins. *J. Biomol. NMR* **1993**, *3* (2), 185−204.

(6) Pons, J. L.; Delsuc, M. A. RESCUE: an Artificial Neural Network Tool for the NMR Spectral Assignment of Proteins. *J. Biomol. NMR* **1999**, *15* (1), 15−26.

(7) Labudde, D.; Leitner, D.; Kruger, M.; Oschkinat, H. Prediction Algorithm for Amino Acid Types with Their Secondary Structure in Proteins (PLATON) Using Chemical Shifts. *J. Biomol. NMR* **2003**, *25* (1), 41−53.

(8) Auguin, D.; Catherinot, V.; Malliavin, T. E.; Pons, J. L.; Delsuc, M. A. Superposition of Chemical Shifts in NMR Spectra Can Be Overcome to Determine Automatically the Structure of a Protein. *Spectroscopy* **2003**, *17* (0), 559−568.

(9) Marin, A.; Malliavin, T. E.; Nicolas, P.; Delsuc, M. A. From NMR Chemical Shifts to Amino Acid Types: Investigation of the Predictive Power Carried by Nuclei. *J. Biol. NMR* **2003**, *30* (1), 47−60.

(10) Seavey, B. R.; Farr, E. A.; Westler, W. M.; Markley, J. L. A Relational Database for Sequence-Specific Protein NMR Data. *J. Biomol. NMR* **1991**, *1*, 217−236.

(11) Vapnik, V. *Statistical Learning Theory*; Wiley: Chichester, Great Britain, 1998.

(12) Cristianini, N.; Shawe-Taylor, J. *An Introduction to Support Vector Machines*; Cambridge University Press: New York, 2000.

(13) Scholkopf, B.; Burges, C.; Smola A. *Advances in Kernel Methods − Support Vector Learning*; MIT Press: Cambridge, MA, 1998.

(14) Scholkopf, B.; Smola A. *Learning with Kernels*; MIT Press: Cambridge, MA, 2002.

(15) Zhang, H.; Neal, S.; Wishart, D. S. RefDB: a Database of Uniformly Referenced Protein Chemical Shifts. *J. Biomol. NMR* **2003**, *25* (3), 173−95.

(16) Collobert, R.; Bengio, S. SVMTorch: Support Vector Machines for Large-Scale Regression Problems. *J. Mach. Learning Res.* **2001**, *1*, 143−160.

(17) Thorsten, J. Making Large-Scale SVM Learning Practical. In *Advances in Kernel Methods − Support Vector Learning*; MIT Press: Cambridge, MA, 1999; Chapter 11.

(18) Wishart, D.; Sykes, B. D. The 13C Chemical-Shift Index: a Simple Method for the Identification of Protein Secondary Structure Using 13C Chemical-Shift Data. *J. Biomol. NMR* **1994**, *4* (2), 171−80.

(19) Wishart, D. S.; Sykes, B. D. Chemical Shifts as a Tool for Structure Determination. *Methods Enzymol.* **2002**, *239*, 363−392.

(20) Sharma, D.; Rajarathnam, K. 13C NMR Chemical Shifts Can Predict Disulfide Bond Formation. *J. Biomol. NMR* **2000**, *18* (2), 165−171.

(21) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28* (1), 235−42.

(22) Auguin, D.; Barthe, P.; Auge-Senegas, M. T.; Hoh, F.; Noguchi, M.; Roumestand, C. 1H, 15N and 13C Chemical Shift Assignment of the Plekstrin Homology Domain of the Human Protein Kinase B (PKB/Akt). *J. Biomol. NMR* **2003**, *27*, 287−288.

(23) Ponchon, L.; Dumas, C.; Kajava, A. V.; Fesquet, D.; Padilla, A. NMR Solution Structure of Mob1, a Mitotic Exit Network Protein and Its Interaction with a NDR Kinase Peptide. *J. Mol. Biol.* **2004**, *337* (1), 167−182.

(24) Gao, G.; Prutzman, K. C.; King, M. L.; Scheswohl, D. M.; DeRose, E. F.; London, R. E.; Schaller, M. D.; Campbell, S. L. NMR Solution Structure of the Focal Adhesion Targeting Domain of Focal Adhesion Kinase in Complex with a Paxillin Ld Peptide: Evidence for a Two-Site Binding Model. *J. Biol. Chem.* **2004**, *279* (9), 8441−51.

(25) Teriete, P.; Banerji, S.; Noble, M.; Blundell, C. D.; Wright, A. J.; Pickford, A. R.; Lowe, E.; Mahoney, D. J.; Tammi, M. I.; Kahmann, J. D.; Campbell, I. D.; Day, A. J.; Jackson, D. G. Structure of the Regulatory Hyaluronan Binding Domain in the Inflammatory Leukocyte Homing Receptor Cd44. *Mol. Cells* **2004**, *13* (4), 483−96.

(26) Banci, L.; Bertini, I.; Cramaro, F.; Del Conte, R.; Viezzoli, M. S. Solution Structure of Apo Cu, Zn Superoxide Dismutase: Role of Metal Ions in Protein Folding. *Biochemistry* **2003**, *42* (32), 9543−53.

(27) Labesse, G.; Colloc′h, N.; Pothier, J.; Mornon, J. P. P-SEA: a New Efficient Assignment of Secondary Structure from C Alpha Trace of Proteins. *Comput. Appl. Biosci.* **1997**, *13* (3), 291−5.

CI050092H