Supporting Information

Computer Vision – Based Automated Peak Picking Applied to Protein NMR Spectra

Piotr Klukowski[2,†], Michal J. Walczak[1,*,†], Adam Gonczarek[2,*,†], Julien Boudet[1] and Gerhard Wider[1*]

[1] Institute of Molecular Biology and Biophysics, ETH Zurich, Otto-Stern-Weg 5, 8093 Zurich (Switzerland)

[2] Department of Computer Science, Wroclaw University of Technology, 50-370 Wroclaw (Poland)

*To whom correspondence should be addressed.
†These authors contributed equally to this work.

Details on procedures mentioned in the main text

Manual Calibration of the Size of the Bounding Box

The size of the smallest and the largest bounding box is defined in such a way that it closely confines the narrowest and the widest real peak in the spectrum. This box size can be set fully manually or with a wizard in the CV-peak-picker that allows choosing from nine pre-defined bounding box sizes. The manual size is adjusted by sliding knobs or by drawing a rectangular box over the peak in the spectrum.

Calculating the sizes of the Bounding Boxes within Feature Pyramid

We first calculate the ratios between the smallest and the largest bounding boxes widths and heights r_w and r_h, respectively. The intervals $[0,1 - r_\mathrm{w}]$ and $[0,1 - r_\mathrm{h}]$ we divide into k equal parts which determine lengths and widths of intermediate sized bounding boxes. For example, for $r_\mathrm{w} = 0.3$ and $r_\mathrm{h} = 0.4$ the differences are 0.7 and 0.6, respectively. For $k = 2$ the widths and the heights of the possible bounding boxes become 0.3, 0.65, 1 and 0.4, 0.7, 1, respectively. Their combination into all possible pairs results in nine bounding boxes. We denote the total number of different bounding boxes by $K = (k + 1)^2$. Here we use $k = 3$ in each dimension, which leads to $K = 16$.

Rescaling of the Peak within the Bounding Box

Since we use different bounding boxes, we have to ensure that HOG features extracted in these boxes will be comparable to each other, *i.e.* will have the same number of features. To do so, the small image surrounded by the bounding box is set to the default size of 32x32 pixels bicubic interpolation (Keys, 1981). We then partition the resulting image into 4x4 cells, each containing of 8x8 pixels (see Fig. 1).

Calibration of r_0 (Powers, 2011)

1. Different values of r_0 are selected (eq. (6)).
2. The values of *recall* vs. *the false-positive rate* are plotted as in Figure S1B below.
3. For r_0 the value is selected for which *recall* reaches 0.98 or more. The *recall* level of 0.98 was arbitrarily chosen and it means that 98% of the real peaks were actually classified as real peaks.

Calibration of gamma in Gaussian kernel (Powers, 2011)

1. Different values of γ are selected in eq. (8).
2. For all the different γ values the classifier is trained on the training set.
3. The quality of classification with different γ is assessed by F-measures on the validation set.
4. The γ value with the best quality classifier is selected for the final version of the program.

Supplementary Figures and Tables

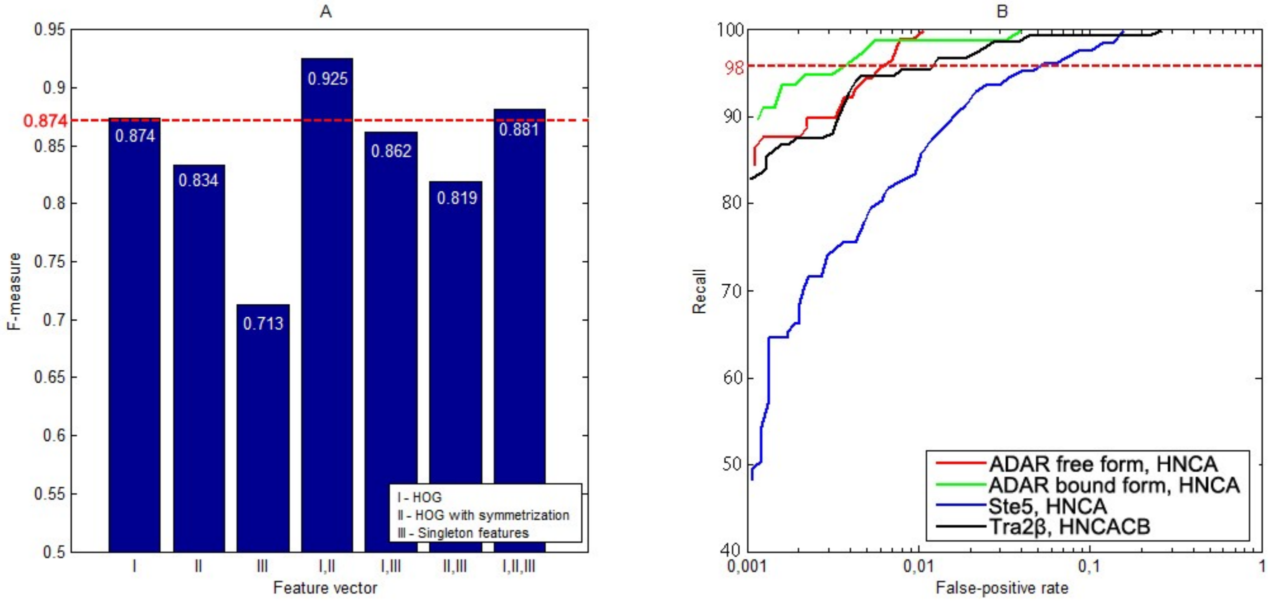


Figure S1. A, bar graph showing the performance (F-measure, Y axis) of the classifiers tested on the testing dataset which consists of NMR spectra of different complexity, size and it contains all type of spectral artifacts. We used HNCA and HNCACB spectra of the following proteins: ADAR (Barraud, et al., 2014), SRSF2 (Daubner, et al., 2012), Tra2β (Cléry, et al., 2011), Ste5 (Walczak, et al., 2014), AF9 (Leach, et al., 2013). The classifiers were trained with different feature descriptors: I - HOG, II - HOG on symmetrized peak, III - set of 13 scalar features (peak intensity, peak volume, peak area, peak width, peak height, width to height ratio, inaccuracy of Gaussian approximation, intensity to height ratio, intensity to width ratio, peak symmetry on horizontal axis, peak symmetry on vertical axis, minimum deviation from the peak center, maximum deviation from peak center). The F-measure level of 0.874 for HOG only (bar labeled I) is indicated by a red dashed line. Best quality of classification is achieved when HOG is calculated before and after peak symmetrization and then both used as one feature vector (bar I,II). B, The Receiver Operating Characteristic (ROC) curves present the classification quality according to classification rule, equation 3, for 4 different triple-resonance spectra (three HNCA and one HNCACB from the testing data set (spectra of ADAR, Ste5 and Tra2β proteins). Recall (Y axis) is defined as TP/(TP+FN), where: TP stands for true positives ("real peaks" which were classified as "true peaks") and FN stands for false negatives ("real peaks" which were classified as "artifacts"); for details see reference 2 (Powers, 2011). Changing the threshold r_0 (eq. (6)), red dashed line at 98, results in an increase or a decrease of the number of peaks selected by the classifier with a concomitant change in the false-positive-rate. An increased *recall* value produces more picked peaks but at the same time a higher false-positive-rate. The red, green, blue and black solid curves represent the peak classification for the spectra as denoted in figure legend.

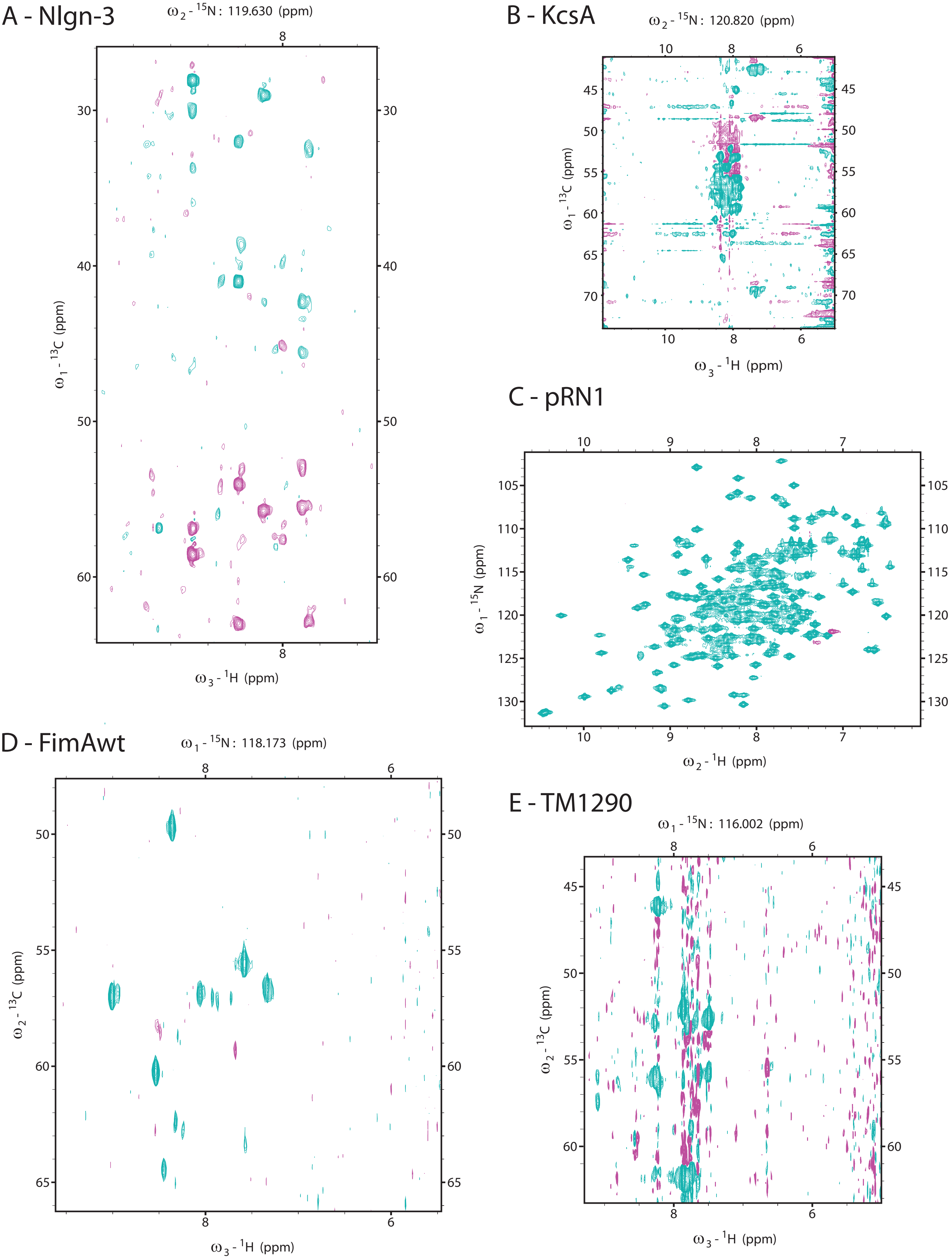
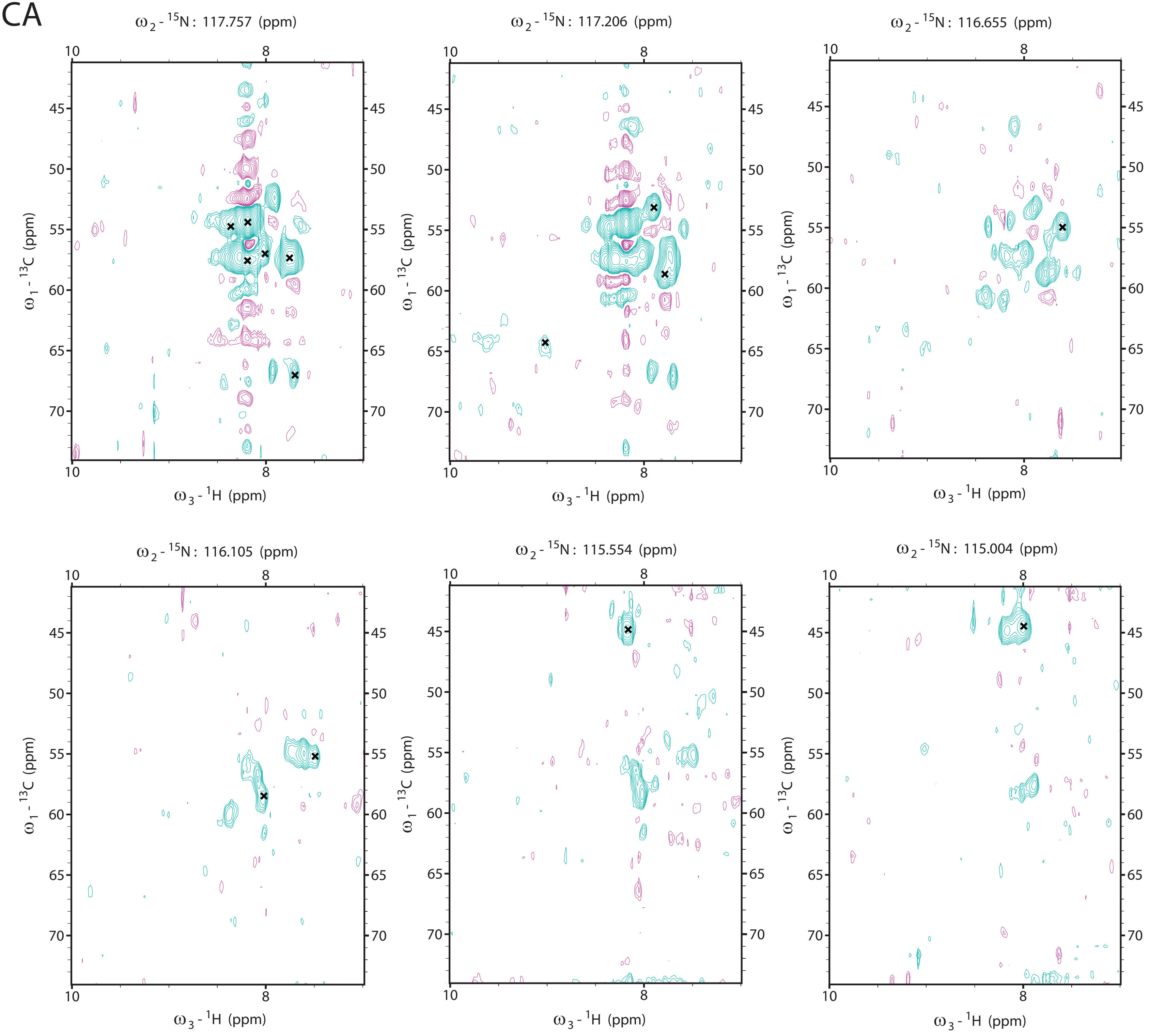

A - Nlgn-3
ω_2 - ^{15}N : 119.630 (ppm)
8
30
40
50
60
ω_1 - ^{13}C (ppm)
ω_3 - ^{1}H (ppm)
B - KcsA
ω_2 - ^{15}N : 120.820 (ppm)
10
8
6
45
50
55
60
65
70
ω_1 - ^{13}C (ppm)
ω_3 - ^{1}H (ppm)
C - pRN1
10
9
8
7
105
110
115
120
125
130
ω_1 - ^{15}N (ppm)
ω_2 - ^{1}H (ppm)
D - FimAwt
ω_1 - ^{15}N : 118.173 (ppm)
8
6
50
55
60
65
ω_2 - ^{13}C (ppm)
ω_3 - ^{1}H (ppm)
E - TM1290
ω_1 - ^{15}N : 116.002 (ppm)
8
6
45
50
55
60
ω_2 - ^{13}C (ppm)
ω_3 - ^{1}H (ppm)

Figure S2. Exemplary spectra used for evaluation of the CV-Peak Picker. A, HNCACB spectrum of Nlgn-3, B, HNCA spectrum of KcsA, C, [^{1}H, ^{15}N] HSQC spectrum of pRN1, D, HNCOCA spectrum of FimAwt and E, HNCA spectrum of TM1290. For all proteins except for pRN1 (2D spectrum) random cross sections along the ^{15}N dimension of the respective 3D spectrum are shown; cyan contours represent positive signals, magenta negative ones. The spectral range between 4.40 and 5.00 ppm in the ^{1}H dimension was excluded from analysis by the program due to heavy distortions by the suppressed water resonance.

KcsA - HNCOCA



ω_2 - ^{15}N : 117.757 (ppm)
ω_2 - ^{15}N : 117.206 (ppm)
ω_2 - ^{15}N : 116.655 (ppm)
ω_2 - ^{15}N : 116.105 (ppm)
ω_2 - ^{15}N : 115.554 (ppm)
ω_2 - ^{15}N : 115.004 (ppm)
ω_1 - ^{13}C (ppm)
ω_3 - ^{1}H (ppm)
10
8
45
50
55
60
65
70

Figure S3. Results obtained by CV-Peak Picker on HNCOCA spectrum of KcsA (Table 2). Six consecutive cross sections (from left to right) along the ^{15}N dimension of the respective 3D spectrum are shown; cyan contours represent positive signals, magenta negative ones; black crosses indicate peaks correctly picked by the program. The correctness was verified manually by experienced NMR spectroscopists. Spectra are shown at noise level and with wide spectral width to present artifacts and water distortions which are not picked by the program. The spectral range between 4.40 and 5.00 ppm in the ^{1}H dimension was excluded from analysis by the program due to heavy distortions by the suppressed water resonance. All real peaks present in the cross sections shown are picked.

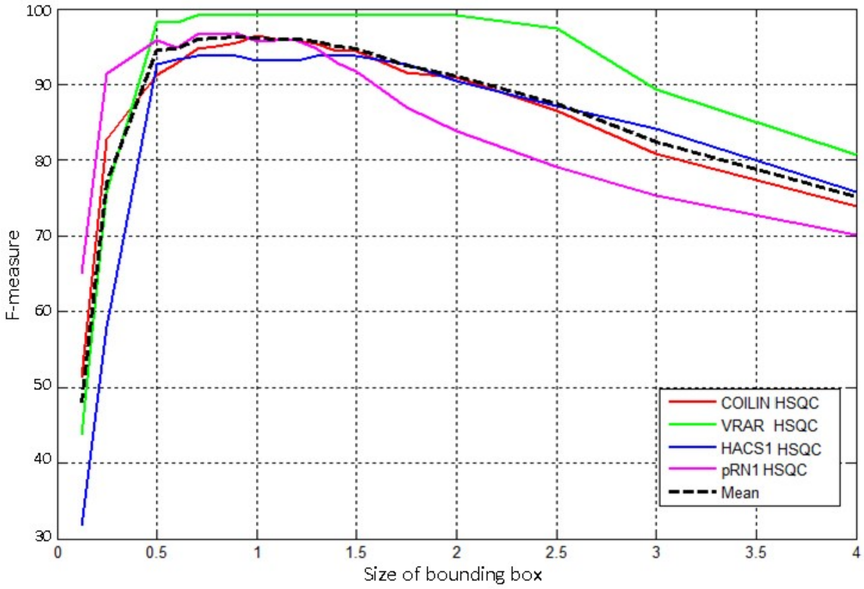


Figure S4. Dependence of the performance (F-measure) of the CV-Peak Picker on the size of the bounding box. The latter is defined as the ratio of the side length of a particular box and the one of the optimal box, i.e., the value 1 represents the optimal bounding box. The curves indicate that CV-Peak Picker maintains its high efficacy when the size of the bounding box is in the range of half to one and a half times the one of the optimal box. The efficacy of the CV-Peak Picker drops dramatically for too small bounding boxes and less pronounced for too large ones.

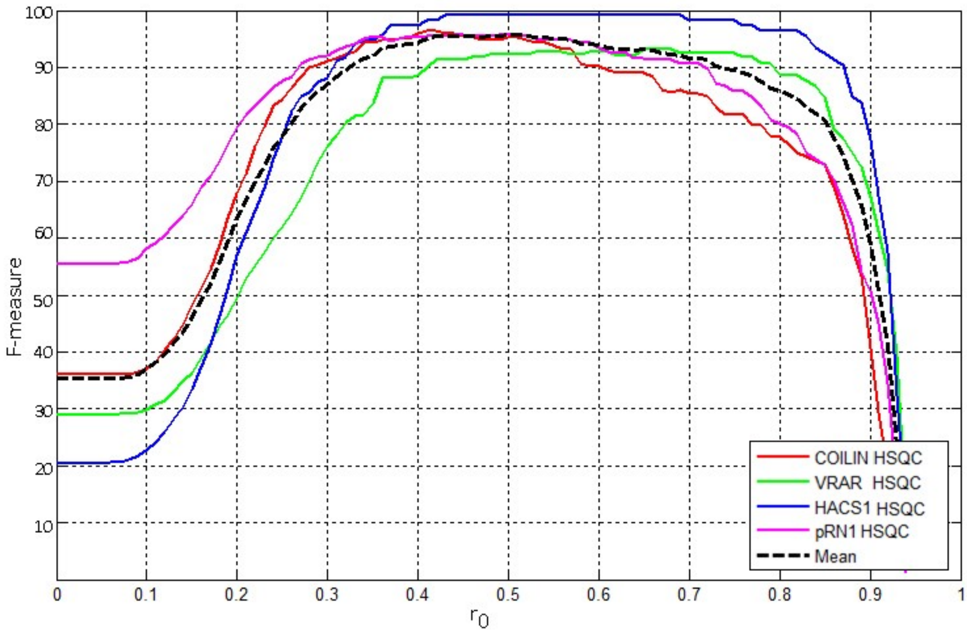


Figure S5 Dependence of the performance (F-measure) of the CV-Peak Picker on the threshold value r_0 (see equ. (6)). The default value of r_0 is 0.5 and in the proximity of this value, the performance of CV-Peak Picker reaches its maximum. The parameter r_0 allows the experienced user a fine tuning of the final peak list.

Table S1. Comparison of the features of different peak picking algorithms(a).

Feature	CV-Peak-Picker	WaVPeak(Liu, et al., 2012)	AUTOPSY(Koradi, et al., 1998)	S. Tikole et. al.(Tikole, et al., 2014)	PICKY(Alipanahi, et al., 2009)
Method for selection of true peaks	Shape-based using Computer Vision (HOG and SVM)(b)	Highest volume for user-defined number of peaks	Peak intensity	Intensities above user-defined threshold	Multistage approach composed of peak pruning, cross-referencing and intensity-based filtering
Deconvolution of overlapping peaks	Symmetrization	None	Segmentation of overlapping peaks and separation based on their symmetry	Decomposition of the overlapped peaks and Noise calculation using factorization of the spectrum with Gaussian kernel	Reconstruction of overlapping peaks using SVD(b) or HOSVD(b).
Noise filtering and initial peak selection	Ranking of peak volumes and initial selection of true peaks candidates (see 'Volume calculation', page 2)	Wavelet smoothing and selection of all extrema in the spectrum	Local noise level estimation based on local variance of spectrum intensities, selection of peaks above local noise level		Assumption of Gaussian noise with variance calculated by comparing intensities of neighboring points in the spectrum

Exclusion of selected signals/regions from analysis	Yes	No	Yes	No	No
Requirement for prior knowledge about the spectrum	No	Number of expected peaks	No	No	No
Extension to other types of molecules and/or spectra	Possible	Possible	Possible	Possible	Possible
Extension to other types of objects in spectra (i.e. artifacts, second conformation etc.)	Possible	Not possible	Not possible	Not possible	Not possible
Graphical User Interface	Yes	No	No	Not specified	No
Underlying NMR processing software	Sparky	Sparky	XEASY(Bartels, et al., 1995)	XEASY	Sparky
Implementation	Java + Matlab	Matlab	ANSI C	Not specified	Not specified
Installation	Not required	Not required	Compilation and installation on the user machine	Not specified	Not required
Special system requirements	Java Virtual Machine, Matlab Environment	Matlab Environment	Not specified	Not specified	Not specified
Platforms	Windows, Linux, Mac	Windows, Linux, Mac	Linux	Linux	Linux
Proteins used for the Evaluation of the peak picker[(c)]	VRAR, HACS1, COILIN, FimAwt, pRN1, KcsA, Nlgn-3, TM1290	VRAR, HACS1, RP3384, CASKIN, TM1112, COILIN, ATC1776, YST0336	WmKT	RcsD-ABL-HPt	VRAR, HACS1, RP3384, CASKIN, TM1112, COILIN, ATC1776, YST0336
Open source	Yes	Yes	No	No	Yes

[(a)] Most cited and newest algorithms were selected.

[(b)] Acronyms used in the table are as follows: HOG – Histogram of Oriented Gradients(Dalal and Triggs, 2005), SVM – Support Vector Machines(Cortes and Vapnik, 1995), SVD – Singular Value Decomposition(Golub and Reinsch, 1970), HOSVD – Higher Order Singular Value Decompostion(De Lathauwer, et al., 2000).

[(c)] The proteins have the following number of amino acids: VRAR - 72, HACS1 - 74, FimAwt - 159, pRN1 - 209, KcsA - 160, Nlgn-3 - 127, TM1290 - 116, RP3384 - 64, CASKIN - 67, TM1112 - 89, COILIN - 98, ATC1776 - 101, YST0336 - 146, WmKT - 88, RcsD-ABL-HPt - 202.

Table S2. Average scanning time T of a 2D layer of a 3D spectrum[a,b].

Protein	Experiment type	T [sec. per layer]
COILIN	CBCA(CO)NH	18.58
	HNCO	19.16
	HNCACB	18.51
	HSQC	20.14
VRAR	CBCA(CO)NH	19.69
	HNCO	21.71
	HNCACB	19.75
	HSQC	21.02
HACS1	CBCA(CO)NH	19.07
	HNCO	19.33
	HNCACB	20.90
	HSQC	23.36
pRN1	HSQC	35.76
KCsA	HN(CO)CA	18.52
FimAwt	HN(CO)CA	20.16
Nlgn-3	HNCACB	19.49
TM1290	HNCA	21.32

[a] Scanning was performed running Matlab on an Intel i7 3632QM processor using 4 threads (“4 Matlab workers”)

[b] Scanning of 500 peaks per layer was set

REFERENCES

Alipanahi, B., *et al.* PICKY: a novel SVD-based NMR spectra peak picking method. *Bioinformatics* 2009;25(12):i268-i275.

Barraud, P., *et al.* A bimodular nuclear localization signal assembled via an extended double-stranded RNA-binding domain acts as an RNA-sensing signal for transportin 1. *Proc Natl Acad Sci U S A* 2014;111(18):E1852-E1861.

Bartels, C., *et al.* The program XEASY for computer-supported NMR spectral analysis of biological macromolecules. *J Biomol NMR* 1995;6(1):1-10.

Cléry, A., *et al.* Molecular basis of purine-rich RNA recognition by the human SR-like protein Tra2-β1. *Nat Struct Mol Biol* 2011;18(4):443-450.

Cortes, C. and Vapnik, V. Support-vector networks. *Machine learning* 1995;20(3):273-297.

Dalal, N. and Triggs, B. Histograms of oriented gradients for human detection. In, *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*. IEEE; 2005. p. 886-893.

Daubner, G.M., *et al.* A syn–anti conformational difference allows SRSF2 to recognize guanines and cytosines equally well. *The EMBO journal* 2012;31(1):162-174.

De Lathauwer, L., De Moor, B. and Vandewalle, J. A multilinear singular value decomposition. *SIAM journal on Matrix Analysis and Applications* 2000;21(4):1253-1278.

Golub, G.H. and Reinsch, C. Singular value decomposition and least squares solutions. *Numerische Mathematik* 1970;14(5):403-420.

Keys, R. Cubic convolution interpolation for digital image processing. *Acoustics, Speech and Signal Processing, IEEE Transactions on* 1981;29(6):1153-1160.

Koradi, R.*, et al.* Automated Peak Picking and Peak Integration in Macromolecular NMR Spectra Using AUTOPSY. *J Magn Reson* 1998;135(2):288-297.

Leach, Benjamin I.*, et al.* Leukemia Fusion Target AF9 Is an Intrinsically Disordered Transcriptional Regulator that Recruits Multiple Partners via Coupled Folding and Binding. *Structure* 2013;21(1):176-183.

Liu, Z.*, et al.* WaVPeak: picking NMR peaks through wavelet-based smoothing and volume-based filtering. *Bioinformatics* 2012;28(7):914-920.

Powers, D.M. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness & correlation. *Journal of Machine Learning Technologies* 2011;2(1):37-63.

Tikole, S.*, et al.* Peak picking NMR spectral data using non-negative matrix factorization. *BMC Bioinformatics* 2014;15(1):46.

Walczak, M.J.*, et al.* The RING Domain of the Scaffold Protein Ste5 Adopts a Molten Globular Character with High Thermal and Chemical Stability. *Angew Chem, Int Ed Engl* 2014;53(5):1320-1323.