

High-Accuracy Peak Picking of Proteomics Data Using Wavelet Techniques

Eva Lange, Clemens Gropl, Knut Reinert, Oliver Kohlbacher, and Andreas Hildebrandt

Pacific Symposium on Biocomputing 11:243-254(2006)

HIGH-ACCURACY PEAK PICKING OF PROTEOMICS DATA USING WAVELET TECHNIQUES*

EVA LANGE[†] CLEMENS GRÖPL, KNUT REINERT
*Institute of Computer Science, Free University of Berlin
Takustr. 9, 14195 Berlin, Germany
E-mail: lange@inf.fu-berlin.de*

OLIVER KOHLBACHER
*Center for Bioinformatics, Eberhard Karls University Tübingen,
Sand 14, 72076 Tübingen, Germany
E-mail: oliver.kohlbacher@uni-tuebingen.de*

ANDREAS HILDEBRANDT
*Center for Bioinformatics, Saarland University,
P.O. 15 11 50, 66041 Saarbrücken, Germany
E-mail: anhi@bioinf.uni-sb.de*

A new peak picking algorithm for the analysis of mass spectrometric (MS) data is presented. It is independent of the underlying machine or ionization method, and is able to resolve highly convoluted and asymmetric signals. The method uses the multiscale nature of spectrometric data by first detecting the mass peaks in the wavelet-transformed signal before a given asymmetric peak function is fitted to the raw data. In an optional third stage, the resulting fit can be further improved using techniques from nonlinear optimization. In contrast to currently established techniques (e.g. SNAP, Apex) our algorithm is able to separate overlapping peaks of multiply charged peptides in ESI-MS data of low resolution. Its improved accuracy with respect to peak positions makes it a valuable preprocessing method for MS-based identification and quantification experiments. The method has been validated on a number of different annotated test cases, where it compares favorably in both runtime and accuracy with currently established techniques. An implementation of the algorithm is freely available in our open source framework OpenMS.

*This work is supported by the German federal ministry of education and research, (grant no. 0312705a 'Berlin Center for Genome Based Bioinformatics').

[†]corresponding author

1. Introduction

Mass spectrometry is one of the central technologies for quantitative proteomics as well as for protein identification. The conversion of the "raw" ion count data acquired by the machine into peak lists for further processing is usually called *peak picking*. This is often done by vendor software bundled with the machine. However, it is often desirable to have more control over this process than one has with the limited intervention allowed by the vendor programs. Any algorithm for peak picking has the following main objectives: First, the peak should have a mass to charge ratio that is as accurate as possible, that means as near as possible to the true mass to charge ratio of the measured compound. This is especially important for identification algorithms. Second, the algorithm should run in real time, that means processing the data should never exceed the time of acquiring it. Among the main difficulties in peak picking are: i) there is often considerable asymmetry in the peaks which confounds a correct mass to charge computation; ii) the convolution of isotopic peaks makes it hard to distinguish individual peaks (this depends on the charge state and the resolution of the instrument). Recently, several approaches to the peak picking problem in proteomics data have been proposed^{1,2,3,4}. For example Strittmatter and coworkers³, use a fit of a Gaussian mixture to model the observed asymmetry in peak shapes. In connection with a calibration method for TOF machines they achieve a considerable improvement in mass accuracy for non convoluted ESI-TOF data. Kempka et al⁴ elaborate on this mixture modelling and test also other mixtures like a Lorentzian and a Gaussian curve. They compare their results to the ones obtained by commercial peak picking algorithms (SNAP) and conclude that they perform better for most peaks. For small and considerably skewed peaks the improvement in accuracy is up to fivefold. The results were obtained on highly resolved MALDI-TOF data without convoluted peaks, since these algorithms require baseline or close to baseline separation of isotopic patterns.

In this paper we describe an algorithm that addresses the above mentioned goals. It computes accurately the mass over charge ratio not only for well-resolved, but also for convoluted data using an asymmetric peak shape. It achieves this in real time and does not make assumptions about the underlying machine or ionization method (MALDI or ESI), which makes the algorithm robust for different experimental settings. This is achieved by addressing the problem from a signal theoretic point of view, which tells

us that spectral data like MS measurements are of an inherently multiscale nature. Different effects, typically localized in different frequency ranges, add up to result in the final signal. In the following, we will assume that the experimentally obtained signal s can be decomposed into three such contributions: a high-frequency noise term n , a low-frequency baseline or background term b , and the information i we are interested in, often referred to as the analytical signal⁵, where i occupies a frequency range in between noise and baseline.

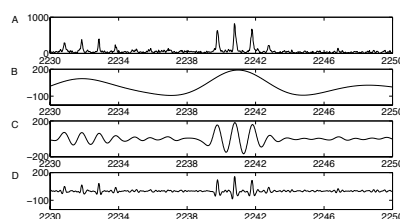


Figure 1. In plot A, B, C, and D the x-axis represents the mass interval between 2230Da and 2250Da, whereas the y-axis shows the intensity. A: Part of a MALDI mass spectrum. Plots B, C, and D show the continuous wavelet transform of the spectrum using a Marr wavelet with different dilation values a (B: $a = 3$, C: $a = 0.3$, D: $a = 0.06$).

The algorithm presented here directly exploits the multiscale nature of the measured spectrum. This becomes possible with the help of a Continuous Wavelet Transform (CWT) – a mathematical tool particularly suited for the processing of data on different scales, where it clearly outperforms classical signal processing techniques like the Fourier Transform, since it preserves information about the localization of different frequencies in the signal in a near-optimal manner⁶. Using the CWT, we can split the signal into different frequency ranges or length scales that can be regarded independently of each other. This is demonstrated in Fig. 1, where we have plotted the transformed signal of a typical region of a mass spectrum on different scales. Apparently, looking at the signal at the correct scale – in our case, a rough estimate of the typical peak width – effectively suppresses both baseline and noise, keeping only the contribution due to the analytical signal. This decomposition allows us to determine each feature of a peak in the domain from which it can be computed best, i.e., either from the frequency range of the analytical signal i , the full signal s , or from a combination of both. Our algorithm is a two-step technique that first determines the positions of putative peaks in the Wavelet-transformed signal

and then fits an analytically given peak function to the data in that region. In an optional third stage, the resulting fit can be further improved using techniques from nonlinear optimization. The method has been validated on a number of different annotated test cases, where it compares favorably in both runtime and accuracy with currently established techniques. The algorithm has been implemented in C++. This implementation is freely available in our open-source framework OpenMS⁷.

In Section 2 we describe the data sets we used and explain our algorithm in more detail. In Section 3 we demonstrate that our algorithm leads to accurate predictions of the mass over charge position and deconvolutes overlapping peaks more accurately than the vendor software. Finally we discuss further developments in Section 4.

2. Methods

2.1. Sample preparation and data generation

Data set A was obtained from a peptide mix (peptide standards mix #P2693 from Sigma Aldrich) of nine known peptides (bradykinin (*F*), bradykinin fragment 1-5 (*B*), Substance P (*H*), [Arg⁸]-vasopressin (*E*), luteinizing hormone releasing hormone bombesin (*G*), leucin enkephalin (*A*), methionine enkephalin (*C*), oxytocin (*D*)). Sample concentration was 0.25 ng/ μ l, injection volume 1.0 μ l. HPLC separation was performed on a capillary column (monolithic polystyrene/-divinylbenzene phase, 60 mm x 0.3 mm) with 0.05% trifluoroacetic acid (TFA) in water (eluent A) and 0.05% TFA in acetonitrile (eluent B). Separation was achieved at a flow of 2.0 μ l/min at 50°C with an isocratic gradient of 0–25% eluent B over 7.5 min. Eluting peptides were detected in a quadrupole ion trap mass spectrometer (Esquire HCT from Bruker, Bremen, Germany) equipped with an electrospray ion source in full scan mode (m/z 500-1500).

Data set B The MALDI-TOF mass spectrum of a tryptic digest of bovine serum albumin (BSA, Aldrich) was acquired from a preparation of an amount corresponding to 50 fmol of the digested protein. In brief, cystines were reduced by incubation with dithiotreitol (DTT) followed by carbamidomethylation using iodoacetamide, prior to proteolysis. The sample was prepared for MALDI using the matrix-affinity sample preparation method with alpha-cyano-4-hydroxycinnamic acid as the matrix⁸. Analysis of positively charged ions in the m/z range 500-5000 was performed on an Ultraflex II LIFT mass spectrometer (Bruker Daltonics, Bremen) operated in the reflectron mode and using Panorama(TM) delayed ion extraction.

A near-neighbour calibration was performed using a peptide standard mixture.

2.2. *The general scheme of our algorithm*

A peak picking technique suitable for high-throughput proteomics applications necessarily needs to combine high accuracy with computational efficiency to provide the results in real time. In our case, a great gain in performance without any negative impact on the accuracy can be achieved by decomposing the mass spectra into smaller parts, so called *boxes*, with a typical length of 10 Da. When splitting a spectrum in this way, special care has to be taken to preserve the signal content. In particular, the split point must not belong to any of the peaks we want to find. After this decomposition of the signal, we compute the wavelet transform of the data in each box. Starting from the maximum position in the wavelet transform, every peak centroid, its height, and its area can be estimated in the raw data. Using these parameters, we are able to represent the raw data peaks using an asymmetric sech^2 and asymmetric Lorentzian function. At this stage of the algorithm, the fitted analytical description is typically in very good correspondence with the experimental signal. To further improve the quality of the fit, the correlation of the resulting peaks with the experimental data can be increased in a subsequent, optional optimization step. This is of particular importance in two cases: first, if neighboring peaks overlap strongly enough that they cannot be fitted well individually, and second, if the resolution of the experimental data is low.

In pseudocode, the general scheme of the algorithm can be formulated as follows:

```
for all mass spectra  $s$  in experiment do
   $box\_list = \text{splitSpectrum}(s)$ 
  for all boxes  $b$  in  $box\_list$  do
     $peak\_list = []$ 
     $w_b = \text{continuousWaveletTransformation}(b)$ 
    while  $\text{getNextMaximumPosition}(w_b, b, x_0)$  do
       $(x_l, x_r) = \text{searchForPeakEndpoints}(b, x_0)$ 
       $c = \text{estimateCentroid}(x_l, x_r)$ 
       $h = \text{intensity}(x_0)$ 
       $(A_l, A_r) = \text{integrateAreas}(x_l, x_r)$ 
       $f = \text{fitPeakShape}(A_l, A_r, c, h)$ 
       $\text{push}(f, peak\_list)$ 
```

```

    removeRawDataPeak( $x_l, x_r, b$ )
     $w_b = \text{continuousWaveletTransformation}(b)$ 
  end while
  optimizePeakParameter( $peakList, b$ )
end for
end for

```

In the following we elaborate on the individual steps.

Decomposing the mass spectrum Our decomposition algorithm ensures that all raw data points belonging to one peak lie in exactly one box. To this end, we split a mass spectrum at a raw data point x if and only if its intensity y is smaller than a given noise threshold, otherwise we search for a minimum in x 's neighborhood using the moving average method. If there is no minimum inside a certain search radius we can be sure that x does not belong to a peak with sensible width, and thus cut at x .

Computing the maximum in the Continuous Wavelet Transform

The Continuous Wavelet Transform $W_{s(b)}$ of the signal s at position b is defined as

$$W_{\psi} s_a(b) = \frac{1}{\sqrt{a}} \int_{-\infty}^{\infty} s(x) \psi\left(\frac{x-b}{a}\right) dx,$$

with the wavelet ψ and the corresponding dilation parameter a . For ψ we choose the so-called Marr wavelet

$$\psi(x) = (1 - x^2) \exp\left(\frac{x^2}{2}\right) = \frac{d^2}{dx^2} \exp\left(-\frac{x^2}{2}\right)$$

since it is known that at least for symmetric peaks, the maximum position in the CWT coincides with the maximum position in the data⁹ and is a good first estimate even for asymmetric peaks. For a peak picking application, the scale of the wavelet (and thus its width) should correspond to the typical width of the peaks. On such a scale, ψ is described by only a few data points and thus the convolution of wavelet and signal can be computed very efficiently with pre-tabulated values of ψ .

Searching for the peak endpoints Defining the “ends” of a peak shape becomes difficult when effects like noise or overlapping of peaks have to be considered. In this case, we cannot expect that the peak's intensity drops below a given threshold before the next peak's area of influence is reached. To solve this problem, we start at the maximum position and proceed to the

left and right until either a minimum is reached, or the value drops below a pre-defined noise threshold. A minimum might either be caused by the rising flank of a neighboring peak, or could be a mere noise effect. To decide between these two cases, we consider again the CWT in the neighborhood, where noise effects are typically smoothed out and peaks can be clearly discerned.

Estimating the peak's centroid For the estimation of the peak's centroid c we compute the intensity-weighted average of the cap of the peak, which is defined as the consecutive set of points next to the maximum with intensity above a certain percentage of the peak's height.

Fitting asymmetric Lorentzian and sech² functions In the literature, several different analytical expressions have been proposed for the representation of mass spectrometrical peaks. Since to our knowledge no universally accepted peak shape exists, our algorithm can fit the data to different peak functions. In the current implementation, we use asymmetric Lorentzian (\mathfrak{L}) or sech² (\mathfrak{S}) functions, which are defined by

$$\mathfrak{L}_{h,\lambda(x),c}(x) = \frac{h}{1 + \lambda^2(x) \cdot (x - c)^2}, \quad (1)$$

and

$$\mathfrak{S}_{h,\lambda(x),c}(x) = \frac{h}{\cosh(\lambda(x)(x - c))^2} \quad (2)$$

where

$$\lambda(x) = \begin{cases} \lambda_l, & x \leq c \\ \lambda_r, & x > c \end{cases} \quad (3)$$

but other peak shapes like double Gaussian profiles^{3,4} can be easily included. A peak can be fitted to the raw data in several ways. In our implementation, we have chosen to use the peak's previously determined centroid and the area under the experimental signal. Fitting the area of the peak automatically introduces a smoothing effect, yields very good approximations to the original peak shape, and is extremely efficient, since the peak's width can be computed from its area in constant time for the functions considered here. Since the peaks are modelled as asymmetric functions, we integrate from the left endpoint x_l up to the peak centroid c to obtain the left peak area A_l . Analogously, we compute the right peak area A_r between c and the right peak endpoint x_r . From these values, we

can finally analytically compute the asymmetric Lorentzian or sech² function with centroid position c and height h which has the same area A_l as the raw peak between x_l and c , and A_r between c and x_r , respectively.

Optimizing the peak parameters The peaks computed so far typically yield a reasonable approximation of the true signal, especially for well resolved, clearly separated peaks. To further improve accuracy, we perform an additional (optional) optimization step. This turned out to be particularly useful for poorly resolved data with strongly overlapping, convoluted peak patterns. Let us assume that we have found m peak functions f_i in a box b , which contains n raw data points (x_i, y_i) , $i = 1..n$. In the previous stage, each of the peaks has been fitted independently of the others, but for a true separation, we need to fit the sum of all peaks to the experimental signal. This can be achieved using standard techniques from nonlinear optimization, like the Levenberg-Marquardt algorithm¹⁰. Each function f_i is described by an initial set of four parameters $p_i = (c_i, h_i, \lambda_{l_i}, \lambda_{r_i})$ with height h_i , centroid c_i , and left/right widths $\lambda_{l_i}/\lambda_{r_i}$. The m parameter sets p_i define the parameter vector $p = (p_1, \dots, p_m)^T$. As loss function, we employ the absolute difference $l_i(p) = |\sum_{j=1}^m f_j(x_i) - y_i|$ between estimated and experimental signal. The nonlinear least squares problem then consists in finding the parameter vector p which minimizes the total loss function

$$\Phi(p) = \frac{1}{2} \sum_{i=1}^n l_i(p)^2 = \frac{1}{2} \|L(p)\|^2 \quad (4)$$

The initial approximation provided by the first stage of the algorithm is usually a good starting point, enhancing the convergence properties. In addition, this allows us to penalize strong deviation from the initial solution, resulting in significantly enhanced robustness.

3. Results

Assessing the quality of a peak picking scheme is a non-trivial problem for which no straight-forward and general approach exists. Obviously, such an algorithm should compute the peak's centroid, height, and area as accurately as possible while featuring a high sensitivity and specificity. To determine the accuracy of, e.g., a peak's centroid, the correct mass value is needed, and thus peak picking algorithms are typically tested against a spectrum of known composition, e.g., a standard peptide mixture or the tryptic digest of a certain protein. Comparing the features of the peaks

found in the spectrum with the theoretical values then gives a measure of the algorithm's capabilities, typically expressed in the average absolute and relative deviation (measured in ppm). Unfortunately, these results are heavily affected by the quality of the experimental data, and additional issues like calibration. Consequently, peak picking algorithms are typically tested against particularly well-resolved spectra, and internal calibration methods are employed. This usually results in high mass measurement accuracy, but the quality of the peak picking algorithms can not be judged independently of the quality of the calibration scheme. From a user's perspective, on the other hand, obtaining similarly well resolved spectra is often infeasible, and internal calibration is not always an option. We have thus decided to demonstrate the capabilities of our approach mainly on LC-MS data with low resolution, containing severely overlapping isotope patterns. Obviously, this complicates comparison of the resulting mass accuracy to published results for alternative peak picking schemes that were tested on well resolved data subjected to sophisticated calibration. We have therefore decided to use the vendor supplied Bruker DataAnalysis 3.2 software on the same spectra to provide a fair means of comparison.

To assess the performance of our peak picking scheme on a set of LC-MS runs on the peptide mixture (dataset A), we determined how often each peptide was found in the expected retention time interval, whether the corresponding isotope patterns (given by at least three consecutive peaks) were discovered and separated, and computed the resulting relative errors of the monoisotopic peak's centroid compared to the theoretical monoisotopic mass. The same analysis was performed with the Bruker software, using the Apex algorithm recommended for ion trap data. The resolution of the data set is critically low with a Δm value of 0.2, implying that each peak is represented by as little as 3–6 data points, and instead of a sophisticated calibration, we only allowed for a constant mass offset to keep the number of fit parameters as small as possible. Using recommended signal-to-noise settings in the Bruker software turned out to miss a large number of the isotopic patterns due to the poor quality of the data. We therefore decided to perform two tests against the Bruker software, one with the recommended setting, and one with a significantly reduced signal-to-noise threshold and peak bound, leading to a total number of peaks comparable to our method. The results of these tests are shown in Table 1. For each peptide, this table contains the theoretical monoisotopic mass, the average relative error of the monoisotopic position, and the number of scans in which the peptide was correctly identified.

Table 1. Evaluation of dataset A. In the table, I denotes the results of the method presented here, II_a the Apex algorithm with reduced thresholds, and II_b Apex with default settings.

<i>z</i>	<i>m</i> _{theo} [Da]	<i>rel. err.</i> [ppm]			<i>#occ.</i>			
		I	II _a	II _b	I	II _a	II _b	
<i>A</i>	1	556.2693	31	35	39	22	35	19
<i>B</i>	1	573.3071	16	24	16	29	57	29
<i>C</i>	1	574.2257	44	60	21	19	44	15
<i>D</i>	1	1007.4365	25	-	94	8	0	5
<i>E</i>	1	1084.4379	18	-	12	3	0	2
<i>F</i>	2	1061.5614	56	64	64	3	2	2
<i>G</i>	2	1183.5730	15	-	-	7	0	0
<i>H</i>	2	1349.736	28	-	13	8	0	1
<i>I</i>	2	1620.8151	37	-	-	13	0	0

Considering the resolution of the raw data, and the lack of sophisticated internal calibration, the mass accuracy that was obtained in these experiments is remarkable. Particularly important is the behaviour on highly convoluted charge two isotopic patterns: as can be seen from the number of correctly identified and separated patterns shown in Table 1, the algorithm presented in this work successfully deconvolutes significantly more of these patterns than the established approaches. The high quality of this separation typically obtained after the optimization stage of our algorithm is shown in Figure 2. In addition, it should be mentioned that the algorithm runs in real time. On the LC-MS spectra of about 100 Mb of data, the peak picking stage took several seconds on a typical PC, while the following optimization run lasted for about 1 to 5 minutes, depending on the number of iterations performed. The applicability of the proposed scheme is not restricted to low-resolution data, nor to ESI data. To demonstrate this, we performed a peak picking on a well-resolved, but difficult MALDI-MS spectrum of a tryptic digest of bovine serum albumin (data set B). This time we performed a Mascot¹¹ peptide mass fingerprinting query with the peaks determined by our implementation and by the Bruker software. In both cases, the bovine serum albumine was identified with a very high significance, where the results obtained with the vendor software led to a sequence coverage of 44% and our peak picking scheme achieved between 52% and 67%, depending on the applied signal-to-noise threshold. It should be noted that for these results, no internal calibration was performed on

the spectrum in order to prevent distortion of the results by possible overfitting due to the calibration procedure. Consequently, the resulting mass accuracy for the peptides identified by Mascot is low with about 95 ppm for Bruker and about 80 to 93 ppm for our method. A simple linear calibration using four monoisotopic masses turned out to reduce the mass error significantly to about 20 to 30 ppm for the same sequence coverages mentioned above.

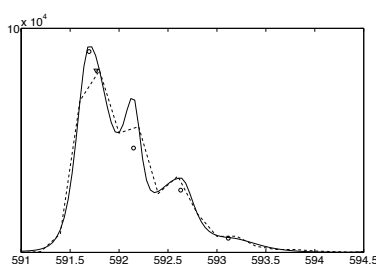


Figure 2. Charge two isotopic pattern of LHRH Decapeptid (solid line: sum of the fitted asymmetric peak shapes, dashed line: linearly interpolated raw data, circles: peak centroids with corresponding peak heights (OpenMS), triangle: peak centroid with corresponding peak height (Bruker Apex)). The relative error of the centroids of the first four peaks as determined by our method are given by 21 ppm, 1.2 ppm, 35 ppm, and 16 ppm.

4. Discussion

We have presented a wavelet-based peak picking technique suited for the application to the different kinds of mass spectrometric data arising in computational proteomics. In contrast to many established approaches to this problem, the algorithm presented here has been particularly designed to work well even on data of low resolution with strongly overlapping peaks. This is especially apparent when deconvoluting for example charge two isotopic patterns with poor separation, as those arising in the LC-MS datasets discussed above. Here, the good performance of our algorithm can be attributed to two of its unique features: the ability to determine the end points of a peak even if it overlaps heavily with another one, which is due to the use of the Wavelet transform as discussed in Section 2, and the optional nonlinear optimization following the peak picking stage. Applied to a high-quality MALDI-TOF spectrum of a tryptic digest, our algorithm yields a high degree of sequence coverage when used as input for a Mascot

fingerprinting query. In all applications, it compares very favorably with the algorithms supplied by the vendor of the mass spectrometers. A free open source implementation is available in the OpenMS C++ framework.

Acknowledgements

We would like to thank Bruker Daltonics, Germany for providing access to the CDAL library for direct access to its raw data formats. In particular, we would like to thank Dr. Jens Decker for his support and helpful discussions. In addition we would like to thank Prof. Christian Huber, Saarland University, Saarbrücken for providing data set A, Dr. Johan Gobom, Max-Planck-Institute for Molecular Genetics, Berlin for providing data set B, and Michael Kerber who was involved in the implementation of an early version of the algorithm.

References

1. Breen, E., Hopwood, F., Williams, K., Wilkins, M. *Electrophoresis* **21** (2000) 2243–2251
2. Yasui, Y., McLerran, D., Adam, B., Winget, M., Thornquist, M., Feng, Z. *Biomed. Biotechnol.* **2003** (2003) 242–248
3. Strittmatter, E.F., Rodriguez, N., Smith, R.D. *Analytical Chemistry* **75** (2003) 460–468
4. Kempka, M., Sjö Dahl, J., Roeraade, J. *Rapid Communications* **18** (2004) 1208–1212
5. Tan, H., Brown, S. *Journal of Chemometrics* **16** (2002) 228–240
6. Louis, A., Maass, D.: *Wavelets: Theory and Applications*. John Wiley & Sons (1997)
7. Kohlbacher, O., Reinert, K.: *OpenMS – an open source framework for shotgun proteomics in C++*. <http://sourceforge.net/projects/open-ms> (2005)
8. Gobom, J., Schuerenberg, M., Mueller, M., Theiss, D., Lehrach, H., Nordhoff, E. *Analytical Chemistry* **73** (20013) 434–438
9. Wu, S., Nie, L., Wang, J., Lin, X., Zhen, L., Rui, L. *Journal of Electroanalytical Chemistry* **508** (2001) 11–27
10. Press, W., Teukolsky, S., W.T., V., Flannery, B.: *Numerical Recipes in C++: The art of scientific computing*. Cambridge University Press (2002)
11. Perkins, D., Pappin, D., Creasy, D., Cottrell, J.: *Probability-based protein identification by searching sequence databases using mass spectrometry data*. *Electrophoresis* **20** (1999) 3551–3567