FEBS Letters 562 (2004) 202–206 FEBS 28187

A new approach that allows identification of intron-split peptides from mass spectrometric data in genomic databases

Jens Allmer[1], Christine Markert, Einar J. Stauber, Michael Hippler[1,*]

Lehrstuhl für Pflanzenphysiologie, Friedrich-Schiller-Universität Jena, Dornburger Str. 159, 07743 Jena, Germany

Received 14 November 2003; revised 7 January 2004; accepted 28 January 2004

First published online 3 March 2004

Edited by Gianni Cesareni

Abstract We present a new approach that allows the identification of intron-split peptides from mass spectrometric data in genomic databases. Our algorithm uses small regions of peptide sequence information which are automatically deduced from de novo amino acid sequence predictions together with the molecular mass information of the precursor ion. The sequence predictions are based on selected collision-induced mass spectrometric fragmentation spectra. Fragments of the predicted amino acid sequence are aligned with each of the six frames of the translated genome and the precursor mass information is used to assemble the corresponding tryptic peptides using the sequence as a matrix. Hereby, intron-split peptides can be gathered and in turn verified by mass spectrometric data interpretation tools such as Sequest.

Key words: Intron–exon structure; Proteomics; Genome data; Mass spectrometry; Search algorithm; *Chlamydomonas reinhardtii*

1. Introduction

With the increasing number of sequenced genomes, the need for new computational approaches is evident. Software tools are needed to cover a broad range of applications. One of the biggest obstacles is the correct identification of intron–exon borders. To identify peptides and proteins from mass spectrometric data several strategies have been developed [1–6]. Up to now, none of these approaches has been able to detect peptides which are split by introns when deduced from genomic DNA sequences. Prediction of intron–exon boundaries for the identification of open reading frames using genomic data is performed by numerous software tools [7–15]. However, those predictions are often erroneous [15–18]. This has especially been outlined in a recent study [19]. An experimental verification of the *Caenorhabditis elegans* genome annotation demonstrated that 50% of the predicted genes (about 4000 genes) needed corrections in their intron–exon structures.

Recently, we analyzed light-harvesting proteins from *Chlamydomonas reinhardtii* in a detailed proteomic study [20]. There we realized that Sequest searches with mass spectrometric data identified several peptides in EST databases which could not be detected in the genomic database from *C. reinhardtii*. One possible explanation for this finding is that these peptides are split by introns when deduced from the genomic sequence. It has been estimated that at least 20–25% [21] of tryptic peptides deduced from genomic databases are split by introns. Our new approach enables identification of these peptides in conjunction with mass spectrometric data interpretation tools such as Sequest and thereby defines intron–exon borders. This approach is related to the sequence tag search algorithm [1,4,22] and uses fragments of amino acid sequences generated by de novo amino acid sequence predictions of tandem mass spectrometry (MS/MS) data together with the corresponding peptide mass of the respective precursor ion. We named this newly developed algorithm the GenomicPeptideFinder (GPF).

2. Materials and methods

2.1. GPF data input

Queries for GPF were generated using de novo amino acid sequencing software (DeNovoX, Thermo Finnigan). Results produced by DeNovoX with a relative probability equal to or greater than 0.1 were queried by GPF. Queries can include monoisotopic or average masses and an identification string for the peptide:

Query e.g.[RZ]AAYPG[VV]CFNPYNLGK

Z represents a cysteine that is carbamidomethylated (plus 57 Da).

2.2. Computer equipment

GPF was originally programmed in Java™ and was tested on several platforms:

1. Pentium II, 400 MHz, 256 MB RAM, Windows 98
2. Pentium III, 966 MHz, 256 MB RAM, Windows XP (Laptop)
3. Pentium IV, 2400 MHz, 1024 MB RAM, Windows 2000
4. IBM RS/6000, F80 4-Way RS64 III 450 MHz Proc, UNIX

2.3. GPF functions and settings

Each amino acid sequence prediction is computationally fragmented (all possible sequence stretches of a given size are produced) and the resulting fragments are used to search for identities in the six frame translation of the genomic database. Two searches are performed: the first one with a longer stretch of amino acids to cut down on processing time and the second one with a shorter sequence which is only invoked if the first one results in matches in one of the

*Corresponding author. Fax: (49)-364-1949232.
E-mail address: m.hippler@uni-jena.de (M. Hippler).

[1] Present address: Department of Biology, University of Pennsylvania, Philadelphia, PA 19104, USA.

Abbreviations: GPF, genomic peptide finder; MS, mass spectrometry; MS/MS, tandem mass spectrometry

0014-5793/04/$30.00 © 2004 Published by Elsevier B.V. on behalf of the Federation of European Biochemical Societies.
doi:10.1016/S0014-5793(04)00212-1

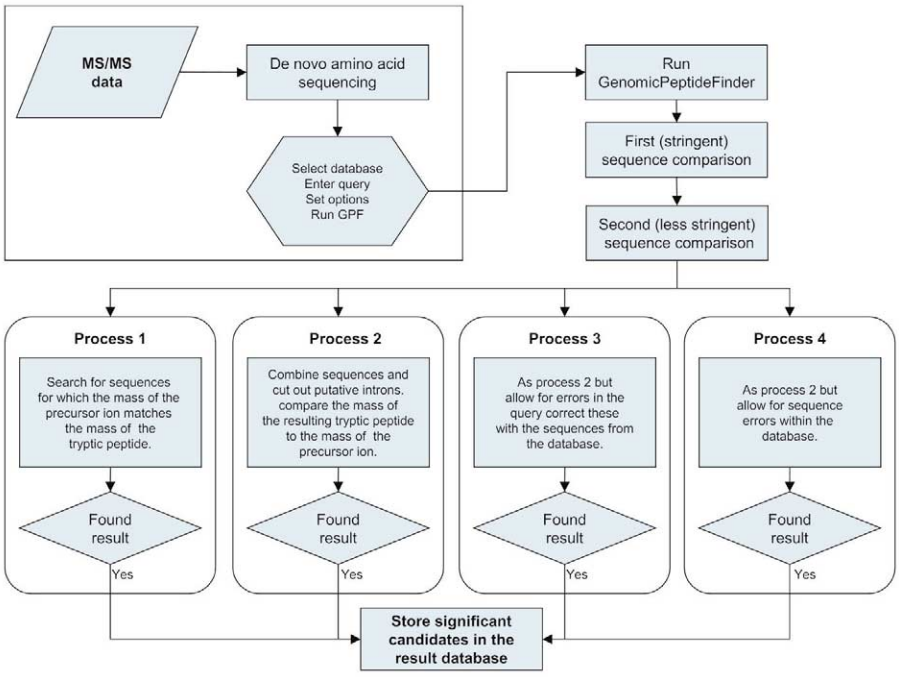

Fig. 1. Abstract and simplified flow chart of the work flow of the GPF. Possible candidates are stored in a result database for further validation.

genomic scaffolds. The more stringently the first comparison is set (a setting of five matching amino acids was used in this study), the faster the search. For the second search, the number of matching amino acids was lowered to three (in this study). In case an amino acid sequence matches the translated genomic sequence (≥ 5 amino acids), the adjacent tryptic cleavage sites are predicted and the respective mass of the deduced tryptic peptide is determined. Then, four different processes are activated that are aimed to identify peptides in a genomic database which match the search criteria. Selected peptides are stored in a database (result database). When the respective precursor mass matches the tryptic peptide's mass, it is extracted and stored in the result database (Process 1, Fig. 1). When process 1 does not result in matching peptides, process 2 is activated which leads to excision of a sequence between two independent amino acid sequence hits in a single scaffold (Process 2, Fig. 1). The respective tryptic peptide is deduced from the genomic data and its mass is matched with that of the precursor ion. In case the masses match within a specified error range, the respective peptide is extracted and stored. When the mass of the tryptic peptide that harbors the two assembled sequence fragments does not agree with the mass of the precursor ion, the sequence is extended from the end of the sequence fragments that hit the translation of the genomic sequence along the corresponding reading frames using the deduced amino acid sequence as a matrix until the resulting tryptic peptide matches the mass of the precursor ion (Process 3, Fig. 1). Process 4 (Fig. 1) operates like process 2 but allows sequence errors in the genomic database. Processes 2–4 lead to the identification of putative tryptic peptides that are split by an intron when deduced from the genomic sequence. For peptide sequences which were found by GPF and were stored in the result database an *E* value and a score were assigned (see below). An additional validation step represents the correlation between the mass spectra from which the de novo predictions and therefore the queries for GPF were originally derived and peptides found by GPF using an MS/MS search tool such as Sequest. These evaluation steps enable GPF to identify peptides which are intron-split within the genomic sequence.

For mass calculations monoisotopic masses were used. The error window for mass deviations between a measured mass and the mass of a deduced tryptic peptide was set at 700 ppm since this is the approximate error window of the ion trap mass spectrometer. A Java™ version of the GPF software with graphical user interface can be obtained for basic research purposes upon request.

2.4. Mass spectrometry and sample preparation

LC-MS/MS analyses were performed with an LCQ Deca XP ion trap mass spectrometer (Thermo Finnigan), which was coupled to a nano-HPLC (Ultimate, LC-Packings), as described [20]. In case the charge state of an ion could not be determined, both doubly and triply charged ion states were taken into account for the de novo amino acid sequence predictions.

Isolation of photosystem I particles and analyses by two-dimensional gel electrophoresis were performed as described in [20].

2.5. De novo predictions

DeNovoX™ (Thermo Finnigan) was used to interpret mass spectra which could not be identified by Sequest. The software provides two markers for prediction quality. One is the absolute probability which is only subject to the assumption that the chemical species being sequenced is a peptide. The other, the relative probability, further assumes that all necessary sequencing information is included in the spectrum. According to the DeNovoX™ manual an absolute probability of 20% or more combined with a relative probability of 75% or more is a strong indication that the sequence or subsequence is correct.

We only examined full length sequence predictions with a relative probability equal to or greater than 10%.

2.6. Using GPF to connect de novo predictions to the original spectra

The GPF software defines a database of possible peptides from de novo predictions. This peptide list has to be checked against the original spectrum to actually identify the correct peptide. This is done by a correlation of the original spectrum with the in silico produced spectra based on this database by Sequest or Mascot. GPF, therefore, provides the link between de novo predictions and Sequest or Mascot evaluation of MS/MS data.

2.7. E value and score calculation

An expectation value (*E* value) and a score were calculated. These two values are used as parameters to cut the overall workload. If peptides are clearly insignificant (*E* value) or if they do not resemble the de novo prediction (score) they are not further processed and thus not stored. Threshold values can be set to exclude results. In our case all results were stored which is also the default setting.

The *E* value reflects the expectation for a peptide to occur randomly within the genome. It is therefore dependent on the size of the ge-

Table 1
Peptides that were gathered by GPF were analyzed by Sequest using the result database and the respective MS/MS spectra

MS/MS spectrum (see Fig. 2)	Query as given by de novo amino acid sequence prediction	Relative probability	Tryptic peptide as deduced from the genomic data	Scaffold	X_{Corr}	E value	Score	Number of hits[a]
A	WLQYSEVLH[AR]	0.272	WLQYSEVIHAR	1152	3.477	210.15	0.848	113
			QYSATRTVLHAR	905	1.523	194.57	0.613	
	[PD]SQYSQVLH[AR]	0.316	QDPSYSQVLLPR	77	1.743	208.62	0.778	323
			LGTWFSSGIAHAR	77	1.434	225.21	0.409	
	[LW]QYSEVLH[AR]	0.410	WLQYSEVIHAR	1152	3.477	210.15	0.848	204
			QYSATRTVLHAR	905	1.523	194.57	0.613	
B	CRGSVN[DE][PL]FK	0.247	NFGSVNEDPIFK	3122	3.930	259.54	0.831	127
			CRGSVNDEPLFK	3122	2.275	227.83	0.831	
	CRGSVNFP[PL]FK	0.363	TCRGSVPPPLPDK	78	1.744	177.23	0.691	200
			CRGSVLPAPPLTR	40	1.064	129.97	0.610	
	CRGSVPFN[PN]FK	0.172	GSMDNNSGGEEGGR	158	1.189	282.12	0.345	248
			SACPCRGCWAASR	476	1.020	186.94	0.317	
C	[RZ]AAYPG[VV]CFNPYNLGK	0.253	GSGDAAYPGGPFFNLFNLGK	1152	4.290	456.18	0.599	144
			VVCLMSAIALPYNGGLRPGV	582	1.396	439.02	0.593	

The X_{Corr} factor for each peptide was determined. Additionally E values and score values were calculated for each peptide. The two hits resulting in the best X_{Corr} factors are listed for each query.
[a]All hits are included, also duplicates which occur in case the de novo prediction gives several alternatives.

nome, the sequence and the mass. The sequence in turn depends on the local amino acid distribution as does the mass. A genome size of 100 Mb was used for E value calculations which corresponds to the size of the *C. reinhardtii* genome (assembly 1) that was downloaded from the Chlamydomonas Genetics Center (http://www.biology.duke.edu/chlamy/). A probability below 0.05 is usually considered significant but for readability reasons $-10\log_{10}$ (probability) was calculated so that E values above 106 can be considered significant [23]. To assess the similarity between queries and their corresponding results we introduced a scoring system. The score value would be 1 for a perfect match and decreases with lower identity. The peptides listed in Table 1 meet this significance criterion.

Since the E value and score calculations are merely used to set limiting values we will not present the algorithms deployed here. The exact formulas and their descriptions can be obtained upon request.

3. Results and discussion

To identify peptides in genomic databases, GPF starts with an alignment of an amino acid sequence that originates from an interpretation of a MS/MS spectrum using a DeNovoX amino acid sequencing software with a respective genomic database (Fig. 2). Such de novo amino acid sequence predictions often result in more than one peptide sequence per MS/MS spectrum. For GPF analysis we restricted the search to those de novo peptide sequence predictions that were given a relative probability equal to or higher than 0.1 (Table 1) as defined by the DeNovoX software.

As proof of principle, three examples are discussed in detail. The functionality of process 1 is illustrated by GPF analyses of de novo sequence predictions evaluated from an MS/MS spectrum that also led to the identification of a peptide by searching the *Chlamydomonas* EST and genomic databases using Sequest [20]. The identified peptide WLQYSEVIHAR is derived from the *lhca3* gene product and is not split by an intron in the nuclear *lhca3* gene present in scaffold 1152 of the genomic database. Three de novo amino acid predictions of that MS/MS spectrum (Table 1; A) were queried by GPF. GPF found the peptide WLQYSEVIHAR two times in scaffold 1152 which is due to several similar de novo prediction alternatives. The E value together with score calculations by GPF for this peptide indicate that this peptide is the most significant tryptic peptide sequence among the sequences found by GPF when analyzing the respective sequence queries. The next two examples derive from MS/MS spectra that did not result in significant Sequest scores using the genomic database but matched when searching the EST *Chlamydomonas* database. Evaluation of the MS/MS spectra shown in Fig. 2B,C by DeNovoX resulted in the prediction of several amino acid sequences that were used as queries for GPF. These queries resulted in numerous tryptic peptide sequences which were found to be significant and good candidates for further evaluation and were therefore collected by GPF. Among these sequences peptide NFGSVNEDPIFK as well as peptide GSGDAAYPGGPFFNLFNLGK resulted in E values in conjunction with scores derived for the queries deduced from MS/MS spectra 1B and 1C, respectively (Table 1), that were very promising. Analysis of the respective MS/MS spectra using the Sequest algorithm and a database that contains the peptides found by GPF can be used to prove that the given peptides are represented by the MS/MS spectra from which the de novo amino acid predictions were obtained. This analysis resulted in cross-correlation factors

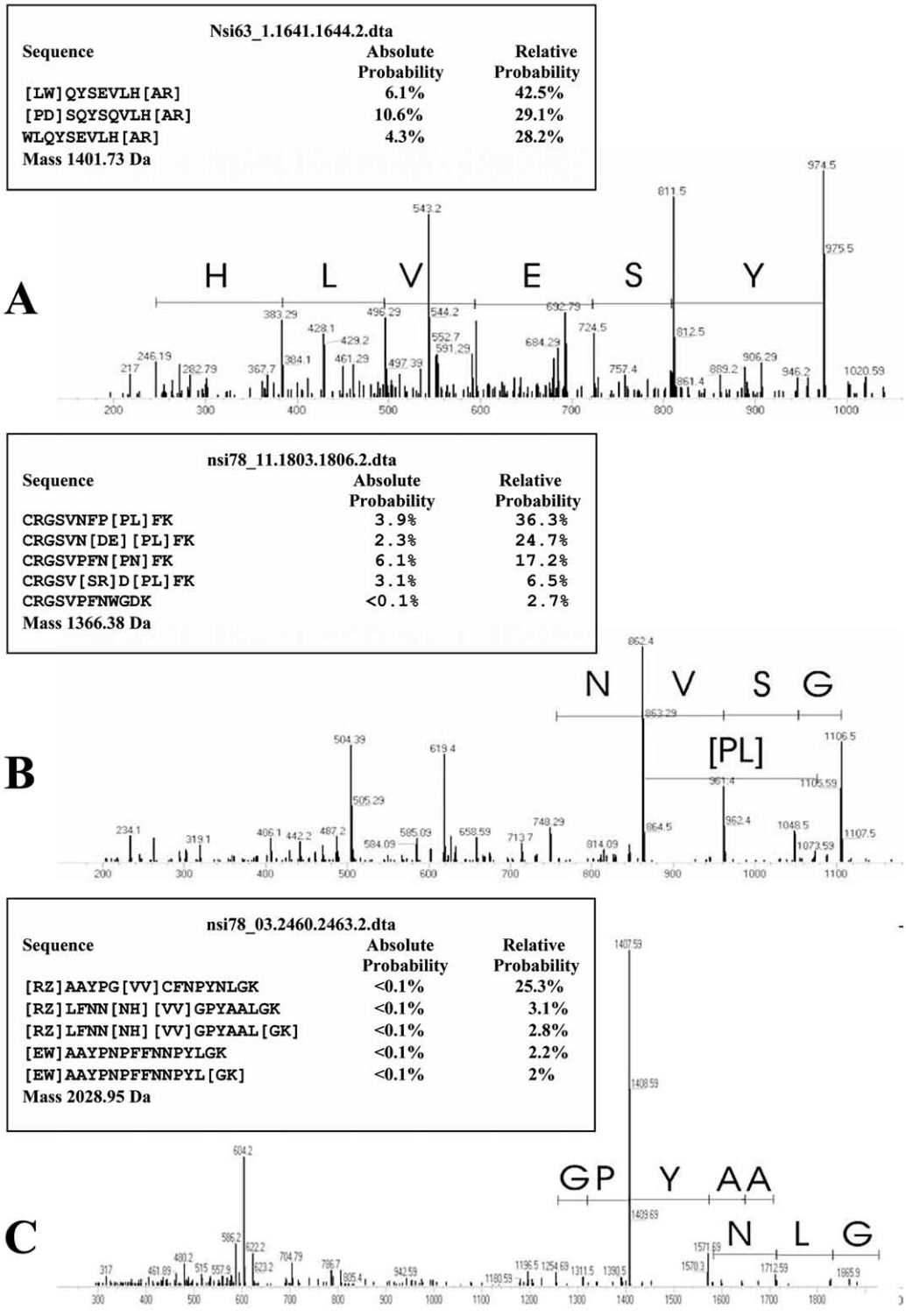

Fig. 2. MS/MS spectra of doubly charged precursor ions and the resulting de novo amino acid sequence predictions. (Z = Cys+57 Da.)

(X_{Corr}) well above 2.5 for the peptides NFGSVNEDPIFK and GSGDAAYPGGPFFNLFNLGK. An X_{Corr} value larger than 2.5 is considered to be significant for fragmentation spectra of doubly charged precursor ions [6] (Table 1). The Sequest search therefore confirmed the GPF predictions. For peptide NFGSVNEDPIFK GPF predicts the following intron–exon boundary: NFGSVNE-intron-DPIFK. This peptide can be deduced from the *lhca5* gene product [20] and is split by an intron in the nuclear gene exactly as predicted by GPF. In this case the amino acid sequence predicted by DeNovoX allowed GPF to determine the correct intron–exon boundary (Fig. 1, Process 2). For peptide GSGDAAYPGGPFFNLFNLGK

GPF matched sequence fragments AAYPG and NLGK. The peptide sequence was extended to the next tryptic cleavage site on the left border of AAYPG that defines the sequence GSGD using the deduced amino acid sequence as a matrix. However, the mass of the resulting peptide GSGDAAYPGNLGK did not match the mass of the precursor ion. Therefore, the peptide sequence was extended on the right border of AAYPG and/or left border of NLGK again using the deduced amino acid sequence as a matrix. Amino acids are added to prolong the peptide sequence until the mass of the peptide matches the mass of the precursor ion or exceeds it, which then terminates the process (Fig. 1, Process 3). Insertion of sequence GPFFNLF resulted in peptide GSGDAAYPGGPFFNLFNLGK which matches the mass of the precursor ion and thus defines an intron–exon boundary within the peptide as GSGDAAYPG-intron-GPFFNLFNLGK. Searching in the *Chlamydomonas* EST database with this peptide sequence revealed that it can be deduced from the *lhca3* gene product [20]. The coding region for this peptide consists of two exons split by an intron exactly as predicted by GPF.

We conclude that our approach enables the identification of peptides which are split by introns in the genome. In addition, our approach has the ability to verify and annotate mistakes in genomic sequences using mass spectrometric data (Process 4, Fig. 1). We suggest that our software tool can be used to complement Sequest or Mascot search tools when mass spectrometric data are used to search in genomic databases to significantly increase the number of identified peptides and proteins.

Acknowledgements: This work has been supported by grants of the Federal State of Thüringen (Nachwuchsgruppe Pflanzenphysiologie) and the Deutsche Forschungsgemeinschaft to M.H.

References

[1] Mann, M. and Wilm, M. (1994) Anal. Chem. 66, 4390–4399.
[2] Giddings, M.C., Shah, A.A., Gesteland, R. and Moore, B. (2003) Proc. Natl. Acad. Sci. USA 100, 20–25.
[3] Mann, M., Hojrup, P. and Roepstorff, P. (1993) Biol. Mass Spectrom. 22, 338–345.
[4] Sunyaev, S., Liska, A.J., Golod, A. and Shevchenko, A. (2003) Anal. Chem. 75, 1307–1315.
[5] Perkins, D.N., Pappin, D.J., Creasy, D.M. and Cottrell, J.S. (1999) Electrophoresis 20, 3551–3567.
[6] Eng, J., McCormack, A.L. and Yates, J.R. (1994) J. Am. Soc. Mass Spectrom. 5, 976–989.
[7] Hutchinson, G.B. and Hayden, M.R. (1992) Nucleic Acids Res. 20, 3453–3462.
[8] Xu, Y., Mural, R., Shah, M. and Uberbacher, E. (1994) Genet. Eng. 16, 241–253.
[9] Burge, C. and Karlin, S. (1997) J. Mol. Biol. 268, 78–94.
[10] Claverie, J.M. (1997) Hum. Mol. Genet. 6, 1735–1744.
[11] Henderson, J., Salzberg, S. and Fasman, K.H. (1997) J. Comput. Biol. 4, 127–141.
[12] Reese, M.G., Eeckman, F.H., Kulp, D. and Haussler, D. (1997) J. Comput. Biol. 4, 311–323.
[13] Pertea, M., Lin, X. and Salzberg, S.L. (2001) Nucleic Acids Res. 29, 1185–1190.
[14] Larsen, T.S. and Krogh, A. (2003) BMC Bioinformatics 4, 21.
[15] Majoros, W.H., Pertea, M., Antonescu, C. and Salzberg, S.L. (2003) Nucleic Acids Res. 31, 3601–3604.
[16] Burset, M. and Guigo, R. (1996) Genomics 34, 353–367.
[17] Guigo, R. (1997) Comput. Chem. 21, 215–222.
[18] Guigo, R., Agarwal, P., Abril, J.F., Burset, M. and Fickett, J.W. (2000) Genome Res. 10, 1631–1642.
[19] Reboul, J. et al. (2003) Nat. Genet. 34, 35–41.
[20] Stauber, E.J., Fink, A., Markert, C., Kruse, O., Johanningmeier, U. and Hippler, M. (2003) Eukaryot. Cell 2, 978–994.
[21] Choudhary, J.S., Blackstock, W.P., Creasy, D.M. and Cottrell, J.S. (2001) Proteomics 1, 651–667.
[22] Shevchenko, A. et al. (1996) Proc. Natl. Acad. Sci. USA 93, 14440–14445.
[23] Perkins, D.N., Pappin, D.J., Creasy, D.M. and Cottrell, J.S. (1999) Electrophoresis 20, 3551–3567.