REVIEW

The pitfalls of proteomics experiments without the correct use of bioinformatics tools

David G. Biron[1], Christine Brun[2], Thierry Lefevre[1], Camille Lebarbenchon[1, 3], Hugh D. Loxdale[4], François Chevenet[1], Jean-Paul Brizard[5] and Frédéric Thomas[1]

[1] GEMI, UMR CNRS/IRD 2724, Centre IRD, Montpellier, France
[2] IBDML, UMR6216, Parc Scientifique de Luminy, Marseille, France
[3] Station Biologique de la Tour du Valat, Le Sambuc, Arles, France
[4] c/o Royal Entomological Society, London, UK
[5] UMR 5096 (UP-IRD-CNRS), Centre IRD, Montpellier, France

Received: March 30, 2006
Revised: June 26, 2006
Accepted: July 10, 2006

The elucidation of the entire genomic sequence of various organisms, from viruses to complex metazoans, most recently man, is undoubtedly the greatest triumph of molecular biology since the discovery of the DNA double helix. Over the past two decades, the focus of molecular biology has gradually moved from genomes to proteomes, the intention being to discover the functions of the genes themselves. The postgenomic era stimulated the development of new techniques (*e.g.* 2-DE and MS) and bioinformatics tools to identify the functions, reactions, interactions and location of the gene products in tissues and/or cells of living organisms. Both 2-DE and MS have been very successfully employed to identify proteins involved in biological phenomena (*e.g.* immunity, cancer, host–parasite interactions, *etc.*), although recently, several papers have emphasised the pitfalls of 2-DE experiments, especially in relation to experimental design, poor statistical treatment and the high rate of 'false positive' results with regard to protein identification. In the light of these perceived problems, we review the advantages and misuses of bioinformatics tools – from realisation of 2-DE gels to the identification of candidate protein spots – and suggest some useful avenues to improve the quality of 2-DE experiments. In addition, we present key steps which, in our view, need to be to taken into consideration during such analyses. Lastly, we present novel biological entities named 'interactomes', and the bioinformatics tools developed to analyse the large protein–protein interaction networks they form, along with several new perspectives of the field.

Keywords:
2-DE / Bioinformatics / Interactome / Misuses / Protein identification

1 Introduction

Determination of the complete genome sequence of an organism has captured the imagination of researchers because the information so gained is expected to reveal, reasonably or unreasonably, the 'key of life', and in an applied sense the understanding of functional genomics may of course have considerable value, for example, in research on human diseases. The term 'genomics' was originally used in 1920 by Winkler to describe the complete set of chromosomes and their associated genes [1]. Three periods can be distinguished with regard to the collective understanding of the DNA molecule and its functionality, which may be loosely termed 'pregenomic', 'genomic' and 'postgenomic'. Aside from primates and insect disease vectors such as mosquitoes, organisms that are now the focus of national or international genomic sequencing efforts include microbes, plants, herbivorous insects (*e.g.* aphids), nematodes, amphi-

Correspondence: David G. Biron, GEMI/UMR CNRS-IRD 2724, IRD, 911 Avenue Agropolis, BP 64501, 34394 Montpellier Cedex 5, France
E-mail: biron@mpl.ird.fr
Fax: +33-4-67-41-62-99

Abbreviations: CCB, colloidal Coomassie blue; **FDR**, false discovery rate; **SN**, silver nitrate

© 2006 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

bians and fishes (see http://www.tigr.org/; http://www.ebl.ac.uk/genomes/; http://www.hgsc.bcm.tmc.edu/projects/aphid/) [2, 3]. Since the beginning of the postgenomic era, the focus of molecular biology gradually moved from genes and genomes to proteins and proteomes and their functionality. Now that several complete genome sequences have been determined, the biggest task in the postgenomic era will be to identify the functions, reactions, interactions and the location of the gene products in tissues and/or the cells of living organisms.

Bioinformatics is an essential part of proteomics research and requires special practical as well as analytical skills for the correct interpretation of results [4–8]. Since the 1990s and following the large scale, worldwide study of proteins in living cells and concomitant attempts to understand the function and regulation of genes, proteomics has become a huge scientific field. Recently, several papers have emphasised the pitfalls of proteomics studies, more especially in relation to 2-DE experimental design, the misuse of statistical tools available with 2-DE softwares, the high rate of 'false positives' for protein identification, and the dangers of cross-species protein misidentification, *i.e.* apparent homologies [8–14]. Thus, in an attempt to clarify the present state of this area and its associated technologies and problems, we review several bioinformatics tools especially developed for 2-DE experiments, from the analysis of 2-DE gels *per se* to protein identification. We summarise the advantages and the misuses of bioinformatics tools. We also discuss the robustness of traditional experimental designs in 2-DE studies, along with present statistical approaches to aid researchers in finding and identifying protein spots which show significant differential expression linked to biological phenomena. Lastly, we present novel biological entities which have emerged from the postgenomic era, namely the 'interactomes', *i.e.* protein–protein interactions within cells and tissues, and the bioinformatics tools developed to analyse the large interaction networks they form, along with several new perspectives of the field.

2 Bioinformatics tools developed for analysis of 2-DE

2.1 Brief history on the creation and use of computer softwares

O'Farrell [15] was the first to develop 2-DE in the mid-1970s to separate complex mixtures of proteins. Since the development of this technique, numerous studies have been performed to enable the extraction of the proteins of any organism, whilst in addition, other studies have attempted to improve the IEF (*i.e.* using IPG strips) and the basic staining methods involved (*e.g.* Coomassie Blue (CB) and silver nitrate (SN) staining) and to develop new ones such as SYPRO Ruby and Lysine tagging (DIGE) [16–22]. During the period encompassing the mid-1970s to the end of the 1990s, 2-DE was used, for example, to construct proteome maps for many species [23], to study the expression of the proteome during the biological development of an organism [24, 25], to reveal the proteome response of an organism to different kinds of treatments or stress [26], and to compare proteome maps between a range of species and/or between populations of the same species [27–34]. By the close of the century, 2-DE databases were created on the Internet for researchers interested in observing the proteome maps of a particular organism at the tissue level and for specific conditions of extraction and separation of proteins (see www.expasy.ch/ch2d/2dindex.html).

Initially during 2-DE studies, protein spot analysis was performed manually and qualitatively, and without the aid of dedicated computer softwares. The presence or absence of spots was converted into a binary matrix so that clustering and correspondence analyses could be performed [35, 36]. Furthermore, the qualitative analysis performed was traditionally summarised in a table giving the number of common protein spots between treatments as well as the specific ones *per* treatment. Because 2-DE is an imperfect technique – due to the distortion of protein patterns caused by polymerisation and running procedure of gels [37, 38] – the need soon became apparent to develop suitable computer softwares to both align and compare gels. The first softwares were designed on the basis of those used by astronomers for nocturnal mapping of stars, one such software being named 'Tycho' in honour of the famous astronomer Tycho Brahe (1546–1601) [39]. At the end of the 20th century, other pioneering softwares like Kepler from Large Scale Biology Corporation (Rockville, USA) and MELANIE from the Swiss Institute of Bioinformatics were developed [40]. Over the past decade, an important number of commercial softwares involving more powerful algorithms and statistical tools than the previous generations of such programs were designed to help researchers deal with the sheer quantity of data produced. Table 1 summarises the commercial softwares presently available and homemade systems for proteomics researchers [41–46]. A series of articles published elsewhere [8, 46–51] compare and contrast the softwares developed for 2-DE analyses.

2.2 Advantages of softwares

Since 2-DE is a powerful separation technique allowing simultaneous resolution of thousands of proteins contained in the proteome of an organism [38, 52], all the associated 2-DE softwares are required to ensure fast and reliable gel comparison [8, 45, 46]. As such, these softwares are now capable of multiple gel analysis, including filtering of 2-DE images, automatic spot detection, normalisation of the volume of each protein spot, and differential and statistical analysis [8, 45, 46].

Such softwares are proving helpful as bioinformatics tools which allow the differential expression of a given proteome (cell, tissue or fluid of an organism) between different

© 2006 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

© 2006 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

Table 1. Softwares and homemade systems currently available for 2-D gel image analysis

Software name or homemade name	Staining method of gels	Statistical tools	Company or University	Year of arrival	Website or E-mail
Commercial softwares					
ImageMaster™ 2D Platinum and MELANIE 6.0	Conventional staining method (*i.e.* CCB, SN, *etc.*) Fluorescent staining methods as SYPRO Ruby Lysine and cysteine tagging (*i.e.* DIGE)	Scatter plots Descriptive statistics of central tendency and dispersion Factor analysis Hierarchical clustering Normal *t*-test Nonparametric tests Wilcoxon–Mann–Whitney Kolmogorov–Smirnov	Amersham Biosciences and GENEBIO	2005	www.2d-gel-analysis.com
DeCyder	Fluorescent staining methods as SYPRO Ruby Lysine and cysteine tagging (*i.e.* DIGE)	Descriptive statistics of central tendency and dispersion Principal component analysis *K*-mean, *k*-medians and hierarchical clustering Discriminant analysis Normal *t*-test ANOVA	Amersham Biosciences	2000	www.amersham biosciences.com/dige
PDQuest™	Conventional staining method (*i.e.* CCB, SN, *etc.*) Fluorescent staining methods as SYPRO Ruby Lysine and cysteine tagging (*i.e.* DIGE)	Descriptive statistics of central tendency and dispersion Factor analysis Hierarchical clustering Normal *t*-test Nonparametric tests Wilcoxon–Mann–Whitney Kolmogorov–Smirnov ANOVA	BioRad	1998	www.bio-rad.com
Proteomweaver™ Professional Entreprise Entreprise PRO	Conventional staining method (*i.e.* CCB, SN, *etc.*) Fluorescent staining methods as SYPRO Ruby Lysine and cysteine tagging (*i.e.* DIGE)	Normal *t*-test Nonparametric tests Wilcoxon–Mann–Whitney Kolmogorov–Smirnov ANOVA	BioRad	2002	www.definiens.com

Table 1. Continued

Software name or homemade name	Staining method of gels	Statistical tools	Company or University	Year of arrival	Website or E-mail
Phoretix™ 2D Expression (PG200TM)	Conventional staining method (*i.e.* CCB, SN, *etc.*)	Descriptive statistics of central tendency and dispersion Normal *t*-test ANOVA	Non Linear Dynamics	1989	www.nonlinear.com
Phoretix 2D Evolution (PG220TM)	Conventional staining method (*i.e.* CCB, SN, *etc.*) Fluorescent staining methods as SYPRO Ruby Lysine and cysteine tagging (*i.e.* DIGE)	Descriptive statistics of central tendency and dispersion Normal *t*-test ANOVA	Non Linear Dynamics	1991	www.nonlinear.com
Progenesis™ (PG420TM)	Conventional staining method (*i.e.* CCB, SN, *etc.*) Fluorescent staining methods as SYPRO Ruby Lysine and cysteine tagging (*i.e.* DIGE)	Descriptive statistics of central tendency and dispersion Normal *t*-test ANOVA	Non Linear Dynamics	2001	www.nonlinear.com
MODAS™	Conventional staining method (*i.e.* CCB, SN, *etc.*) Fluorescent staining methods as SYPRO Ruby Lysine and cysteine tagging (*i.e.* DIGE)	Descriptive statistics of central tendency and dispersion *K*-mean, *k*-medians and hierarchical clustering Normal *t*-test ANOVA Nonparametric tests Kolmogorov-Smirnov Dunnett's test Kruskall–Wallis	Non Linear Dynamics	2000	www.nonlinear.com
GELLAB II+™	Conventional staining method (*i.e.* CCB, SN, *etc.*)	Descriptive statistics of central tendency and dispersion Normal *t*-test	Scanalytics	1999	www.scanalytics.com
ProteinMine™	Conventional staining method (*i.e.* CCB, SN, etc.)	Descriptive statistics of central tendency and dispersion	BioImagene	2005	www.bioimagene.com
Z3*TM*	Conventional staining method (*i.e.* CCB, SN, *etc.*)	Descriptive statistics of central tendency and dispersion Hierarchical clustering Normal *t*-test	Compugen	2000	www.2dgels.com

© 2006 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

Table 1. Continued

Software name or homemade name	Staining method of gels	Statistical tools	Company or University	Year of arrival	Website or E-mail
AlphaMatch 2-D™	Conventional staining method (*i.e.* CCB, SN, *etc.*)	Descriptive statistics of central tendency and dispersion Normal *t*-test ANOVA	Alpha Innotech Corporation	1999	www.adpsa.co.za
Investigator HT ANALYZER™	Conventional staining method (*i.e.* CCB, SN, *etc.*)	Descriptive statistics of central tendency and dispersion Normal *t*-test ANOVA	Genomic Solutions	2000	www.genomicsolutions.com
Homemade systems					
Fuzzy logics	Conventional staining method (*i.e.* CCB, SN, *etc.*)	Comparison of fuzzy 2-DE maps by means of multivariate statistical tools	University of Eastern Piedmont; University of Verona	2003	righetti@sci.univr.it
Three-way PCA	Conventional staining method (*i.e.* CCB, SN, *etc.*)	Comparison of 2-DE gels (i.e. control and treated samples) by means of three-way PCA	University of Eastern Piedmont; University of Verona	2003	righetti@sci.univr.it
PCA coupled to classification methods and cluster analysis	Conventional staining method (*i.e.* CCB, SN, *etc.*)	Principal component analysis Cluster analysis	University of Eastern Piedmont; University of Verona	2004	righetti@sci.univr.it
Intelligent data mining architecture	Conventional staining method (*i.e.* CCB, SN, *etc.*)	–	University of Sunderland	2006	james.malone@sunderland.ac.uk

© 2006 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

treatments and/or between populations, the aim usually being to find and characterise proteins linked to particular biological phenomena. They thus permit alignment and comparison of 2-DE gels in experiments designed to detect the qualitative and/or quantitative difference between replicates of the same group of samples and between different classes of gels. In 2-DE software outputs, a group represents the same protein spot as displayed in several gel runs, whilst a class of gels represents a number of gels having a common biological meaning or characteristic(s).

2.3 Misuses of softwares

There are various criteria which proteomics researchers need to take into consideration when preparing a 2-DE experiment. These include: (i) the choice of the biological compartment (cell/tissue/fluid); (ii) the method of extraction of proteins; (iii) parameters for the separation of the protein; and (iv), staining methods. Traditionally, colloidal Coomassie blue (CCB) was the most widely used dye; however, it is less sensitive than SN for protein detection (for a review, see [18]).

The choice of the staining method is an important step in any 2-DE experiment [18, 22]. For instance, SN displays excellent sensitivity as a stain and reveals a greater and often more important number of protein spots on a 2-DE gel than CCB for the same amount of soluble protein of an organism used. Consequently, many researchers have used SN staining in 2-DE experiments as their method of choice to reveal the maximum number of protein spots. However, some studies have reported that the advantage of using SN for higher detection sensitivity compared with CCB is counterbalanced by inferior sequence coverage with MS; thus abundant proteins can be expected to yield roughly 11–34% sequence coverage with SN dye, whilst CCB will typically achieve values of 30–67% [8, 53]. Which ever stain is used, 2-DE softwares are thereafter employed to analyse the differential expression between groups and classes of gels. In the case of SN staining, such softwares can be inappropriately used since the linear dynamic range of SN is very weak [18, 22] and indeed, this stain is only truly suitable for the qualitative analysis of results [8, 22]. In addition, use of SN staining in 2-DE experiments may result in a copious number of 'false negatives' for the candidate spots if the quantitative analysis tools provided with the 2-DE softwares are used. Because of this, for any given 2-DE experiment, the implications of choice concerning the staining method need to be assessed before any differential analysis with 2-DE softwares is performed. For example, for quantitative analysis many recent fluorescent staining methods have a robust linear dynamic range and are designed to detect the quantitative differential expression between spots in a 2-DE experiment [18, 22]. As such, the detection of protein with abundances as low as 300 copies *per* cell has been reported [54]. Fluorescent staining methods can be divided into two groups: covalently bound (*e.g.* Alexa-dyes™ and CyDyes™) and noncovalently bound (*e.g.* SYPRO Ruby™ and Deep Purple™). Both groups of fluorescent dyes are linear over at least three orders of magnitude with sensitivity as good as with silver staining [55–57]. Because of this, the use of fluorescent dyes has drastically increased the sensitivity and reproducibility of protein quantification in 2-DE experiments (for more details, see [8, 18, 22]).

Whilst softwares for 2-DE provide for automatic detection of protein spots, normalisation of the volume of each protein spot detected, and differential analysis of protein spots between treatments, some studies reveal that manual intervention is necessary to correct the step of detection of protein spots (*i.e.* deletion of false protein spots and correction of the shape of protein spots) as well as the pairing step of protein spots within a same class (category) of gels and between different classes (categories) of gels [8]. Such intervention is thus essential to prevent a number of false positives for the candidate protein spots.

As an example of this, in one of our recent studies we exposed a hairworm species, *Paragordius tricuspidatus* (Dufour) (Nematomorpha, Chordodidae), to four treatments (control plus three biological treatments) [58]. The gels were stained with SN. We then compared the number of protein spots obtained for each category following an automatic analysis (Image Master 2D Platinum Software Version 5.0) with those detected using a semiautomatic analysis (*i.e.* manual intervention to verify the steps of detection of protein spots and of pairing of protein spots between treatments). Table 2 reveals there to be a significant difference in the total number of protein spots (specific + common protein spots) observed between the two types of analyses

Table 2. Number of protein spots detected by automatic and semiautomatic analyses

Pattern of protein spots	Treatment where protein spots occurred				Number of protein spots	
	Control	Treatments			Automatic analysis	Semiautomatic analysis
		1	2	3		
Common	X	X	X	X	170	358
Specific		X	X	X	1	10
	X				65	14
		X		X	30	21
	X		X		22	6
		X			169	24
	X		X	X	30	5
		X	X		20	10
	X			X	43	7
			X		181	18
	X	X		X	92	36
			X	X	12	3
	X	X			48	14
				X	140	17
	X	X	X		48	10
Total of gel (common + specific)					1071	553

© 2006 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

(χ^2_{14} = 873.49, p <0.0001). As suggested by other authors, manual intervention is still necessary for the differential analysis using these softwares [8, 45, 59–62].

Astonishingly, 2-DE commercial softwares offer the Student's *t*-test (*i.e.* a univariate statistical test) as an approach to detect significant alteration in protein expression in data obtained between two treatments (*i.e.* it is used to determine whether there is a significant difference between the average volumes of the same protein spot made under two different conditions (*i.e.* control and treatment) [63, 64]). Both measurements are made on each unit in a sample, and the test is based on the paired differences between these two average volumes. The null hypothesis is that there is no difference between the mean (μ) values (*i.e.* H_0: $\mu_1 = \mu_2$), and with the null hypothesis being tested against the alternative hypothesis (*i.e.* H_1: $\mu_1 \neq \mu_2$).

When using this test, two major assumptions are that the dataset for each treatment follows a normal distribution and that there are more than three replicates *per* treatment. Testing of normal distribution of 2-DE data has only been mentioned in a handful of proteomics papers. As a matter of fact, two types of distributions need to be evaluated: the distributions of spot volumes of individual spots across replicate gels as well as the distribution of the relative spot volume variances in the replicate gels [8]. Recently, the normality of 2-DE data produced by DIGE and the DeCyder™ software package was evaluated [13]. It was found that: (i) approximatively 95% of spot volume distribution was normal; (ii) variance distribution was non-normal; (iii) log-transformation of data generally used for DIGE data leads to inflated variance at low signal levels; and (iv), arsinh transformation of data is better to normalise DIGE data [13]. Since DIGE data exhibit similar characteristics to microarray datasets, some statistical methods (*i.e.* normalisation of data and adjusting of the *p*-values) developed for microarray analysis were adapted for the DIGE [65].

Because the bioinformatics analysis is traditionally time consuming in 2-DE experiments, only a few replicates (*i.e.* 3 to $\leq$7) tend to be made. As a consequence, the individual variability of protein spots accounted for in this statistical test is poorly estimated with so few replications. For any researcher interested in using the Student's *t*-test available on 2-DE softwares, we suggest a minimum of five replicates *per* treatment [8, 63, 66] and the following steps: (i) testing the normal distribution of protein spot volumes of individual spots across replicate gels of an identical treatment using the Shapiro–Wilk's test [67]; (ii) in the case of data showing a non-normal distribution, use of an arsinh transformation of the 'difference gel electrophoresis' (DIGE) data or log transforming it [61, 68, 69]; and (iii) a *p*-level of at least 0.01 to reduce the number of false candidate protein spots observed.

In the case of data showing a non-normal distribution after transformation, two nonparametric tests can be applied, *viz*. the Wilcoxon–Mann–Whitney and Kolmogorov–Smirnov tests. The Wilcoxon–Mann–Whitney test is one of the most powerful of such tests for comparing two populations. It is used to test the null hypothesis that both populations have identical distribution functions *versus* the alternative hypothesis that the two distribution functions differ only with respect to location (median), if at all (for more details see [63, 64]). For a single sample of data, the Kolmogorov–Smirnov test is used to test whether the data sample is consistent with a specified distribution function: when there are two data samples, it is used to test whether these two samples may reasonably be assumed to come from the same distribution (for more details see [63, 64]).

Furthermore, 2-DE commercial softwares offer One-Way ANOVA as an approach to detect significant alteration in protein expression in data obtained between two or more treatments. This approach thus allows comparison of several groups of observations, all of which are independent but possibly with a different mean for each group. A test of great importance is whether all the means are equal [63, 64]. The null hypothesis here is that there is no difference between the mean (μ) values (*i.e.* H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_k$, where k is the number of treatments), and with the null hypothesis being tested against the alternative hypothesis (*i.e.* H_1: $\mu_1 \neq \mu_2 \neq \mu_3 \neq \mu_k$).

When using this test, two major assumptions are that (i) the dataset for each treatment follows a normal distribution and (ii) there are more than three replicates *per* treatment. In contrast, with the Student's *t*-test, only a few replicates (*i.e.* 3 $\leq$ 7) tend to be made since the bioinformatics analysis is time consuming. For any researcher interested in using the One-Way ANOVA available on many 2-DE softwares, we suggest a minimum of five replicates *per* treatment and the same steps as suggested above for the Student's *t*-test, namely: (i) testing the normal distribution data; (ii) transformation of data if it is necessary to do so; and (iii) a *p*-level of at least 0.01 to reduce the number of false candidate protein spots observed.

Few 2-DE softwares (such as DeCyder from Amersham Biosciences) offer Two-Way ANOVA as an approach to detect significant alteration in protein expression in data obtained between multiple (2 plus) treatments when studying the effects of two factors separately (their main effects) as well as together (their interaction effect). The Two-Way ANOVA is an extension to the One-Way test [63, 64] and has two independent variables (hence the name) called factors. The idea is that there are two factors (variables) which affect the dependent variable (*i.e.* volumes of protein spots detected). Each factor will have two or more levels within it, and the degrees of freedom for each factor is one less than the number of levels. There are three sets of hypotheses with Two-Way ANOVA. The null hypotheses for each of the sets are: (i) the population means of the first factor are equal; (ii) the population means of the second factor are equal; (iii) there is no interaction between the two factors. Five major assumptions to respect when using the test are: (i) the dataset for each treatment follows a normal distribution; (ii) the treatments must be independent; (iii) the variance of the treat-

© 2006 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

ments must be equal; (iv) the treatment groups must have the same sample size; and (v) more than three replicates *per* treatment are used [63, 64].

Overall, an inappropriate utilisation of the Student's *t*-test, One-Way and Two-Way ANOVA may result in an important numbers of false positives [8, 58], which is effectively the case in more than 60% of recent proteomics studies [58]. Moreover, assuming a 5% level of error (*i.e.* $p = 0.05$) for 1000 protein spots means that there are potentially some 50 (*i.e.* 0.05×1000) false positives, which is unacceptable. Multiple testing correction methods, such as the Bonferroni correction [64] and false discovery rate (FDR) [70], adjusts the Student's *t*-test or ANOVA values for each protein spot to keep the overall error rate as low as possible. In the Bonferroni correction, the unadjusted *p*-values are multiplied by the total number of tests performed. The FDR is a less stringent correction method but more practical approach than the Bonferonni correction. The FDR is defined as V/R for $R > 0$ (where V denotes the number of falsely rejected hypotheses and R indicates the total number of rejected hypotheses) and FDR = 0 if $R = 0$. Since V is unobserved, a sequential *p*-values procedure has been developed to control the expected value of the FDR (*i.e.* E(FDR)) under the assumption that the test statistics are independent [70]. The resulting process controls E(FDR) at the fixed level α for any joint distribution of the *p*-values.

2.4 Suggestion of experimental design for 2-DE

Since the original introduction of proteomics approaches, the poor experimental design of 2-DE experiments has tended to be all too commonplace due to the apparent technical difficulties, the high cost of data acquisition as well as the time needed for data analysis [14, 66]. Some recent papers emphasise that a significant number of studies involving 2-DE were done with a nonrigorous experimental design, more especially in relation to the number of replicated gels *per* treatment and the inappropriate application of statistical tests available in 2-DE software packages [8, 14, 45, 62, 66]. Like many other proteomics researchers, we performed few replicates *per* treatment in our earlier 2-DE experiments. Is it reasonable to continue in this way? One major goal of 2-DE experiments is to find protein spots for use as biomarkers which show significant differential expression between different treatments, to understand biological phenomenon, and also to study the proteome variability between populations [14, 38, 45, 52, 71–73]. A new attitude is essential to improve the reliability of proteomics data, both in terms of recording of the protein spots *per se* as well as differential analysis of these.

Clearly during any given proteomics study, experimental design should be improved when and wherever possible, especially regarding the number of replicates *per* treatment, in order to reduce the number of false positive protein spots detected. In addition, whilst the currently available proteomics techniques and bioinformatics tools are powerful means by which to generate high quality data, both in terms of quality and quantity, nevertheless as outlined above, the inappropriate use of statistical tests and involving false assumptions could lead to poor analysis of the data collected and worse still, to a false understanding of the biological process(es) investigated. The rigorous application of statistics and with due cognisance to the assumptions being made, will to some extent increase the work done during such 2-DE experiments, but on the other hand and most importantly, it will improve the reliability of the results achieved.

Some researchers outline the different steps essential for proteomics study of such biological phenomena as host–parasite interactions [71–76], biological development [77], toxicology [78, 79] and the response of an organism's genome to environmental stress [2]. Figure 1 outlines some key steps that need to be taken into consideration for the realisation of a 2-DE experiment (*i.e.* from the number of replications *per* treatment to the appropriate statistical analysis according to the choice of staining method).

3 Protein identification and bioinformatics

3.1 Protein identification and imperative need of softwares

In any given proteomics study, a key step during the research is the identification of proteins linked to the biological phenomenon under investigation, whatever this may be exactly [5, 52, 80–82]. This imperative has stimulated the development of many new instruments [83–90] and techniques [90, 91], especially protein databases [92] and bioinformatics tools [7, 93, 94].

Edman's method was one of the first techniques to be developed for the identification of proteins [95, 96]. However, this method is actually considered too slow and expensive for 2-DE studies. Furthermore, proteins are frequently N-terminally blocked. By the 1990s, the pitfalls associated with this method stimulated the development of new techniques. The quality of 2-DE experiments was improved with the concomitant development of MS instrumentation techniques [83–86].

MALDI-TOF-MS is presently the most popular instrument used for protein identification, *i.e. via* determination of peptide molecular weight [97]. Thus many protein databases and bioinformatics tools have been developed for analysis of the PMF data collected (see Table 3). Over the years, several studies have stressed the pitfalls of PMF for protein identification [14, 98], and with such awareness has come the drive to develop new more powerful techniques and softwares in order to improve such identification. Table 3 summarises the free softwares available on the Internet for protein identification according to the instrument types used to reveal a property or a combination of properties which are unique to

© 2006 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

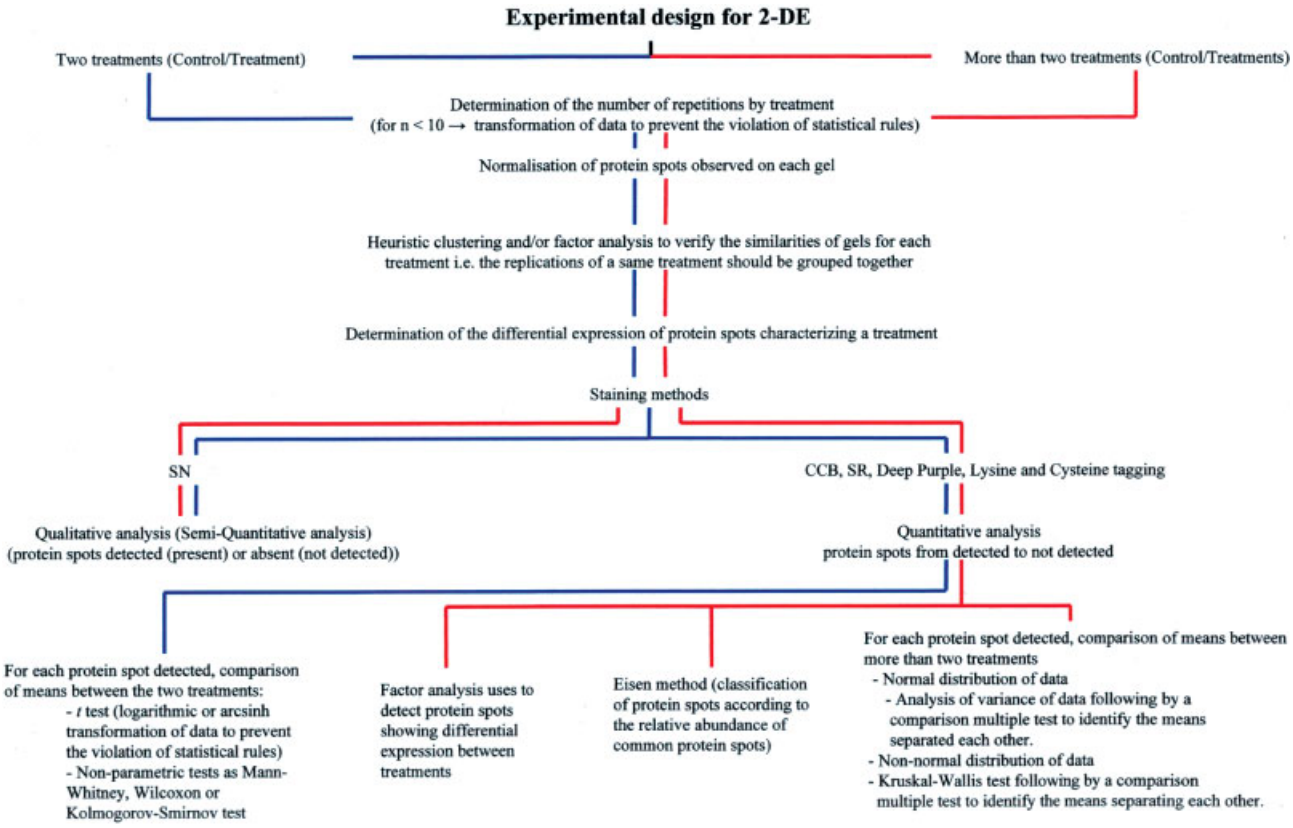


Figure 1. Key steps in the design of a 2-DE experiment staining methods: CCB, colloidal Coomassie blue; SN, silver nitrate; SR, SYPRO Ruby; DP, Deep purple.

the candidate proteins in question and which can be subsequently employed to search databases for sequence match(es).

3.2 Advantages of softwares

2-DE experiments generate a large quantity of potentially important data, data that may require many weeks to analyse and interpret. Since the 1990s, a number of important bioinformatics tools were developed to aid protein identification [4, 5, 7, 45, 71, 72, 99–103]. These softwares permit comparison and/or matching of the observed data (*i.e.* candidate protein spots) with theoretical data from protein databases. A high throughput in search databases with high efficiency and at low cost are two of the major advantages of the current softwares available for protein identification on the Internet (see Table 3).

The new generation of these softwares takes into account many criteria during searches of the currently available protein databases. For protein identification with PMF and MS/MS data, it is possible to specify some spot properties (*i.e.* p*I*, MW and taxon of the organism under study), the protease used (*i.e.* trypsine or others), the number of missed cleavages, the mass spectrometer type and its accuracy, the mass type, the possible amino acid modifications, and more interestingly, the *p*-value threshold to identify a positive match between observed and theoretical protein data [45, 99, 103]. Following the identification of a candidate protein spot, many softwares such as Aldente (http://www.expasy.ch/tools/aldente/), Phenyx (http://www.phenyx-ms.com/) and MASCOT (http://www.matrixscience.com/) allow access to an impressive number of crossreferences *via* different databases, *i.e.* InterPro (http://www.ebi.ac.uk/interpro/, PANTHER (http://www.pantherdb.org/), PFAM (http://www.sanger.ac.uk/Software/Pfam/), PRINTS (http://umber.sbs.man.ac.uk/dbbrowser/PRINTS/), *etc.*) as well as to different softwares, allowing, for instance, the modelling of the quaternary structure of the identified protein by using the Swiss-Model (http://swissmodel.expasy.org//SWISS-MODEL.html). However, the various softwares available for protein identification do not offer the same advantages. It is not the aim of this review to compare available softwares, but Table 3 can hopefully aid researchers in their choice.

3.3 Misuse of softwares

In proteomics studies, the classic criteria used to confirm protein identification are the MOWSE (molecular weight search) score, *p*-value, % coverage, and the Δppm (difference

© 2006 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

Table 3. List of free protein identification sofwares available on the Internet (June 2006)

Method of identification	Software name	Parameters available for the search on protein databases	Protein databases available	Web link
PMF				
	Aldente	p*I*, MW, taxon, digestion, modifications (variable and/or fixed), missed cleavages, thresholds of spectrometer, peptide scoring	Swiss-Prot, TrEMBL	http://www.expasy.org/tools/aldente/
	MASCOT	MW, taxon, digestion, modifications (variable and/or fixed), missed cleavages, thresholds of spectrometer, peptide scoring	MSDB, NCBI, Swiss-Prot, Random	http://www.matrixscience.com/
	MS-Fit	p*I*, MW, taxon, digestion, modifications (variable and/or fixed), missed cleavages, thresholds of spectrometer, peptide scoring	Genpept, Ludwignr, NCBI, Owl, Swiss-Prot, EST_mouse, EST_human, EST_others	http://prospector.ucsf.edu/ucsfhtml4.0/msfit.htm
	PepMAPPER	p*I*, Mw, taxon, digestion, modifications (variable and/or fixed), missed cleavages, thresholds of spectrometer, peptide scoring	Swiss-Prot, PDB	http://wolf.bms.umist.ac.uk/mapper/
	PeptideSearch	MW, taxon, digestion, modifications (variable and/or fixed), missed cleavages, thresholds of spectrometer, peptide scoring	NRDB	http://www.narrador.embl-heidelberg.de/GroupPages/Homepage.html
	ProFound	MW, taxon, digestion, modifications (variable and/or fixed), missed cleavages, thresholds of spectrometer, peptide scoring	NCBI	http://prowl.rockefeller.edu/profound_bin/WebProFound.exe
MS/MS				
	Popitam	Taxon, digestion, modifications (variable and/or fixed), missed cleavages, thresholds of spectrometer, type of spectrometer, peptide scoring	Swiss-Prot, TrEMBL	http://www.expasy.org/tools/popitam/
	Phenyx	Taxon, digestion, modifications (variable and/or fixed), missed cleavages, thresholds of spectrometer, type of spectrometer, peptide scoring	Swiss-Prot, TrEMBL, MSDB, NCBI, EST_mouse, EST_HUMAN, EST_RAT	http://www.phenyx-ms.com/
	MASCOT	Taxon, digestion, modifications (variable and/or fixed), missed cleavages, thresholds of spectrometer, type of spectrometer, peptide scoring	MSDB, NCBI, Swiss-Prot, Random, EST_mouse, EST_human, EST_others	http://www.matrixscience.com/
	OMSSA	Taxon, digestion, modifications (variable and/or fixed), missed cleavages, thresholds of spectrometer, type of spectrometer, peptide scoring	NCBI	http://pubchem.ncbi.nlm.nih.gov/omssa/

© 2006 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

Table 3. Continued

Method of identification	Software name	Parameters available for the search on protein databases	Protein databases available	Web link
	PepFrag	Taxon, digestion, modifications (variable and/or fixed), missed cleavages, thresholds of spectrometer, type of spectrometer, peptide scoring	NCBI, dbEST	http://prowl.rockefeller.edu/
	MS-Tag	MW, taxon, digestion, modifications (variable and/or fixed), missed cleavages, thresholds of spectrometer, type of spectrometer, peptide scoring	Genpept, Ludwignr, NCBI, Owl, Swiss-Prot, EST_mouse, EST_human, EST_others	http://prospector.ucsf.edu/ucsfhtml4.0/mstagfd.htm
	SearchXLinks	Digestion, modifications (variable and/or fixed), missed cleavages, thresholds of spectrometer, peptide scoring	MSDB, NCBI, Swiss-Prot, Random	http://www.searchxlinks.de/
	PeptideSearch	MW, taxon, digestion, modifications (variable and/or fixed), missed cleavages, thresholds of spectrometer, peptide scoring	NRDB	http://www.narrador.embl-heidelberg.de/GroupPages/Homepage.html
Sequence tags				
	FASTS/FASTF	Number of peptides, score table, filter, descriptions, alignments	UniProt, UniRef100, UniRef90, UniRef50, UniParc, Swiss-Prot, EuroPatents, JapanPatents, USPTO Patents	http://fasta.bioch.virginia.edu/
	MS-Seq	p*I*, MW, taxon, digestion, modifications (variable and/or fixed), AA composition, Instrument	Genpept, Ludwignr, NCBI, Owl, Swiss-Prot, EST_mouse, EST_human, EST_others	http://prospector.ucsf.edu/ucsfhtml4.0/msseq.htm
	MS-BLAST	Number of peptides, score table, filter, descriptions, alignments, other advanced options	NRDB95, sp_NRDB, Swiss-Prot, hs_swiss, PDB, ENSEMBLE PEP	http://dove.embl-heidelberg.de/Blast2/msblast.html
Amino acid sequence (microsequences)				
	MS-Pattern	p*I*, MW, taxon, digestion, modifications (variable and/or fixed)	Genpept, Ludwignr, NCBI, Owl, Swiss-Prot, EST_mouse, EST_human, EST_others	http://prospector.ucsf.edu/ucsfhtml4.0/mspattern.htm
	PeptideSearch	MW, taxon	NRDB	http://www.narrador.embl-heidelberg.de/GroupPages/Homepage.html
Amino acid composition				
	AACompIdent	p*I*, MW, AA composition, calibration protein	Swiss-Prot, TrEMBL	http://us.expasy.org/tools/aacomp/

© 2006 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

Table 3. Continued

Method of identification	Software name	Parameters available for the search on protein databases	Protein databases available	Web link
p*I* and MW and/or sequence tags				
	TagIdent	p*I*, MW, taxon	Swiss-Prot, TrEMBL	http://www.expasy.org/tools/tagident.html
p*I*, MW, PMF, sequence tags, amino acid composition				
	MultiIdent	p*I*, MW, taxon, AA composition, PMF (digestion, modifications, thresholds of spectrometer	Swiss-Prot, TrEMBL	http://www.expasy.org/tools/multiident/

in mass between experimental and theoretical peptides). The question may then be posed, 'is it enough to prevent a mismatch and to link, without error, a protein spot on a gel to a known protein in a database(s)'? Unfortunately, many recent studies and some workshops clearly underline the risk of false positive identifications, more especially for the PMF but also for the MS/MS [14, 45, 52, 94, 98]. Thus, although the MOWSE score and the *p*-value obtained with classic criteria constitute a useful guide in protein identification, they can never be a substitute for the careful interpretive analysis necessary to detect a false from a true positive result [14, 45, 98]. Some reasons for failing to match MS data especially for the MS/MS or the LC-MS/MS when searching a protein database are as follows: (i) the peptide sequence is not in databases; (ii) an unsuspected PTM; (iii) the peptide is a result of nonspecific cleavage; and (iv) the product-ion data is of poor quality [104].

The % coverage (*i.e.* the proportion of a theoretical protein which is covered by MS data of a protein spot) can be used as a parameter by the protein identification softwares such as Aldente (http://www.expasy.org/tools/aldente/) to determine the scoring to find the best match. In most softwares, the % coverage is given mainly as a criterion to help proteomics researchers to find the best theoretical protein matching with the observed MS data.

What is a good coverage? According to many proteomics researchers, more than 20% coverage is likely to be significant [45, 80, 103]. Even so, this criterion is not enough to avoid the risk of false positive identifications. For an organism with a complete genome sequence available in the databases, there is still a high probability of obtaining false positive matches with theoretical proteins with a MW$\geq$40 kDa when a search is done for PMF data for a protein spot of 30–40 kDa without restriction in terms of p*I* and MW criteria [14, 98, 105]. This potential pitfall can be explained in part by the mass redundancy (*i.e.* the fact that peptides with the same amino acid sequence but with different alignments can have the same mass) [45, 52, 89]. Another factor to explain this pitfall is the higher number of theoretical peptides for a protein with a MW greater than 40 kDa. In this case, the probability of an observed peptide matching a theoretical peptide in a database(s) increases with the MW of the molecule concerned.

The risk of misuse of the % coverage for protein identification is very important for an organism with incomplete genome sequences in databases since the search is limited to the nearest species with partial or complete genome sequences. The main hypothesis of cross-species identification is that orthologous proteins share a similar function, and a similar structure and amino acid composition resulting in a sharing of many peptide masses [6, 45, 80, 106]. As some studies emphasise, at least 70% of sequence identity between proteins is necessary for a conservation of the peptides involved [45, 107]. Thus cross-species identification needs to be done with due care and attention [72, 73, 107, 108].

What is the maximum % coverage for a protein match with PMF data? For tryptic digests, the range of experimental mass values is 800–3000 Da. To obtain 100% coverage for a theoretical protein in the databases, the peptide mass values should therefore be in the range of 800–3000 Da, because of its estimation by MS measurement [45, 52, 71, 72, 83]. In general, the theoretical proteins in databases have many peptide mass values higher than 3000 Da and lower than 800 Da. For instance, % coverage of many proteins, whatever the species studied, will never exceed 40% with very good PMF data for protein with a MW range of 30–50 kDa. The NADH dehydrogenase subunit 5 (Fragment) of the butterfly, *Parnassius ruckbeili ruckbeili* (Deckert) (TrEMBL accession number: Q76JY5) has an important number of theoretical peptides with mass values $\gg$3000 Da. Concerning the p*I* and MW, both criteria will be relatively close to theoretical values for species with complete genome sequences in databases, and as far as cross-species identification is concerned, and in

© 2006 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

order to avoid false positive identification, a molecular mass variation of ±30% and a p*I* variation of ±2.0 are generally used [6, 45, 80, 81, 91, 103, 108].

The risk of obtaining false positive identifications is not limited to MS techniques. For instance, it is also true for the Edman method followed by BLASTP searching. Figure 2 gives an example with a protozoan species, *Leishmania major* Friedlin. The complete genome sequences of this protozoan are available in the NCBI database. The complete sequence of a *L. major* protein is presented in Fig. 2A. Three BLASTP were done for this protein: (i) for the complete protein sequences available (Fig. 2B); (ii) for all available protozoan genomes (Fig. 2C); and (iii) for all taxa available in the databases (Fig. 2D). Interestingly and importantly, the BLASTP for the genome of *L. major* did not provide good protein identification. The second BLASTP on protozoan genomes recognised that it is a protein from the NUDIX family, and only with the third BLASTP for all taxa that the protein was

A)

```
  1 mkhtyvrtgl evvsglkftr lcsllytatd dgaqpgnkwe mvqrttrsta lsaferspap
 61 ipvdaveica vvrrsskrfi vvvaqyrppv dsvclefpag lvddnenagq aairemheet
121 gfvvdetdiv sispplstep gltdsccvlv rldvdgerae nqkpkqhldd gedievllip
181 isqpknalna lsdvvkryae kgqraivdak lytfmealaw gv
```

B)

Search with BLASTP with Leishmania major genome

Sequences producing significant alignments:	*Score (Bits)*	*E Value*
ref|XP_848156.1| FtsJ cell division protein [Leishmania major st	28.9	0.24
ref|XP_848033.1| hypothetical protein LMJ_0336 [Leishmania ma...	27.7	0.53
ref|XP_848050.1| hypothetical protein LMJ_0353 [Leishmania ma...	25.8	2.0
ref|XP_843177.1| Leishmania major strain Friedlin hypothetica...	24.6	4.5

C)

Search with BLASTP with Protozoa genomes

Sequences producing significant alignments:	*Score (Bits)*	*E Value*
ref|XP_829045.1| NUDIX hydrolase [Trypanosoma brucei TREU927]	208	7e-54
gb|EAN93969.1| nudix hydrolase, putative [Trypanosoma cruzi]	197	2e-50
gb|EAN89729.1| nudix hydrolase, putative [Trypanosoma cruzi]	197	2e-50
ref|XP_640150.1| hypothetical protein DDB0204862 [Dictyosteli...	98.6	1e-20
gb|EAL45665.1| ADP-sugar pyrophosphatase, putative [Entamoeba...	85.9	7e-17
gb|EAN32108.1| hypothetical protein TP04_0755 [Theileria parva]	70.9	2e-12
gb|EAA37795.1| GLP_549_31342_32085 [Giardia lamblia ATCC 50803]	47.8	2e-05
ref|XP_645669.1| hypothetical protein DDB0216797 [Dictyosteli...	40.4	0.004
gb|EAL46026.1| mutT/nudix family protein [Entamoeba histolytica	36.6	0.051
gb|EAN30785.1| hypothetical protein TP03_0049 [Theileria parva]	32.7	0.73

D)

Search in BLASTP for all Taxon

Sequences producing significant alignments:	*Score (Bits)*	*E Value*
gi|68130023|emb|CAJ09331.1| nudix hydrolase-like protein [Leishm	443	2e-123
gi|74024958|ref|XP_829045.1| NUDIX hydrolase [Trypanosoma bru...	208	1e-52
gi|71654400|ref|XP_815820.1| nudix hydrolase [Trypanosoma cru...	197	3e-49
gi|71420709|ref|XP_811580.1| nudix hydrolase [Trypanosoma cru...	197	3e-49
gi|50798650|ref|XP_424023.1| PREDICTED: similar to ADP-sugar ...	124	2e-27
gi|39592916|emb|CAE62530.1| Hypothetical protein CBG06639 [Caeno	118	1e-25
gi|76660924|ref|XP_880835.1| PREDICTED: similar to ADP-sugar ...	113	6e-24
gi|17564994|ref|NP_503726.1| NuDiX family member (ndx-2) [Cae...	112	8e-24
gi|73949060|ref|XP_535190.2| PREDICTED: similar to ADP-sugar ...	112	1e-23
gi|55962510|emb|CAI11760.1| novel protein (zgc:86930) [Danio ...	111	2e-23

Figure 2. (A) Complete protein sequence of nudix hydrolase-like protein of *Leishmania major* (AN in NCBI (gi|68130023|emb|CAJ09331.1|)). Results of BLASTP for gi|68130023|emb|CAJ09331.1| with *L. major* genome (B), protozoan genome (C) and all taxa (D).

© 2006 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

duly recognised. This example further illustrates that considerable attention and interpretive skills are necessary whatever the techniques and bioinformatics tools used for protein identification.

Some pitfalls for BLASTP were revealed in previous studies [109, 110], but in the case of *L. major* proteins as described above, this is the first example of a pitfall in relation to a species with a completely sequenced genome. Thus, to avoid this pitfall, we suggest making the protein identification in three steps when using the taxon field of the BLASTP software, *viz.*: (i) at the species level, identification of the protein and its family; (ii) at the family or order level, confirmation of the protein family of the candidate protein; and (iii), confirmation of the identification of the candidate protein without restriction in the taxon field (*i.e.* all organisms).

3.4 Recommendations for identifying candidate proteins obtained by 2-DE

Several methods are routinely used to identify proteins from 2-DE experiments. These methods rely on comparisons with sequence databases derived from genomic programmes, cDNA studies, protein sequencing, ESTs or genomic sequence tags (GSTs). MS is a core approach in proteomics. One MS approach, the PMF with limited exceptions, cannot be applied to short stretches of sequences such as ESTs. But the MS/MS or LC-MS/MS approaches can be used in synergy with ESTs databases to identify proteins of species with unsequenced genomes [104, 111] and rapid characterisation of a protein mixture [112].

As protein identification is without doubt a key step in 2-DE experiments, it is important to use certain 'tricks of the trade' to reduce the possibility of obtaining and misinterpreting false positive results. Figure 3 summarises some key steps in the process of protein identification for species with complete and incomplete genome sequences in databases. In addition to these key steps, proteomics researchers should use as many properties of the candidate protein as possible (p*I*, MW, *etc.*) to increase the probability of obtaining a match in terms of the 'best' theoretical protein data available [6, 45, 83, 95].

Protein sequence databases are growing at a near-exponential rate. Whilst this is generating massive amounts of information for some model species as *Drosophila* and humans, the majority of species remain more or less molecularly undefined. The magnitude of this is illustrated in the Swiss-Prot database, where 38% of all sequences derive from just ten organisms [91]. Can the data from species with completed genome sequences be used for the identification

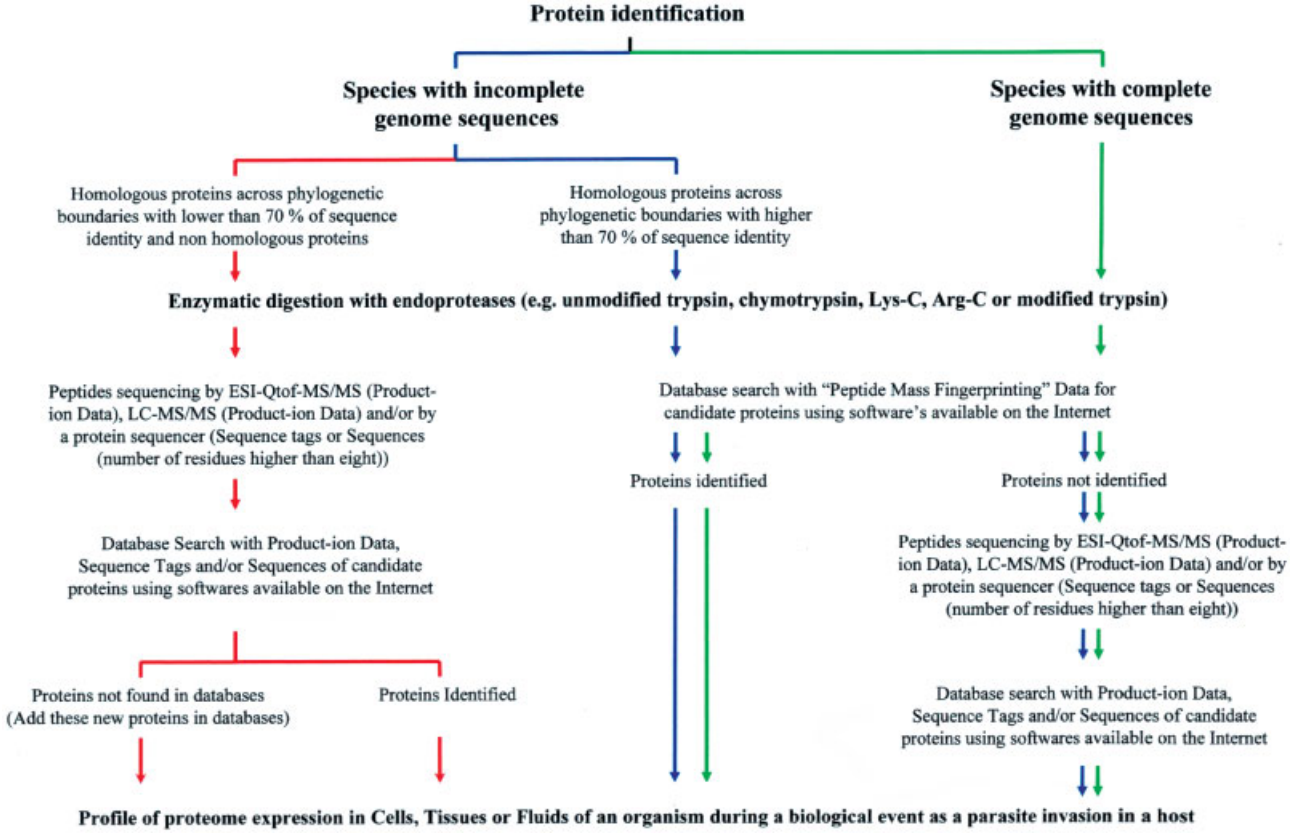


Figure 3. Key steps to take into consideration in the process of protein identification.

© 2006 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

of proteins from other species with uncompleted genome sequences? An answer to this question is that the search for cross-species identification should be limited to the nearest taxon to the species studied and the interpretation should be conservative. This is because some examples are known where similar proteins have very different functions in different species [113], whilst by way of contrast, very different proteins have similar functions [114]. Cross-species identification requires a careful usage of bioinformatics tools along with very careful interpretation of the methods employed for peptide/protein identification [45, 80, 91, 103, 115].

A theoretical study of 65 cross-species comparisons involving 21 different types of protein has revealed some clear findings in relation to the attributes of these proteins in cross-species protein conservation [72]. The p*I* was found to be poorly conserved with some proteins showing as much as p*I* ± 2 difference across species boundaries. On comparison, the intact mass of proteins was well conserved, with a mean absolute difference of 1.9%. Peptide masses were not well conserved across species boundaries, with few or no peptides being conserved when sequence identity between two proteins was lower than 70% [45, 72]. However, amino acid composition of proteins was well conserved across species boundaries, with many proteins failing to show large compositional differences between species until sequence identity was $\ll$60%. The poor conservation of peptide mass data is expected, as a single amino acid substitution in any peptide can drastically change its mass. The MultiIdent tool (http://www.expasy.org/tools/multiident/) from Expasy is one of the best softwares for cross-species identification since it can accept many attributes of proteins: (i) p*I*; (ii) mass; (iii) composition; (iii) peptide masses (PMF data); (iv) sequence tags; and (v) choice of the taxon (*i.e.* species, genus, family or kingdom levels).

3.5 Data representation

Researchers often use mathematical clustering methods to reveal interesting patterns in large datasets, such as those produced by proteome analysis following protein identification. Users then need interactive visualisation tools to facilitate pattern extraction, identifying for instance proteins with similar profiles and thus possibly with similar functions [116]. Many softwares are available on the Internet for graphic visualisation as THEA (http://thea.unice.fr/summary-en.html) and TreeDyn (http://www.treedyn.org/).

THEA is an integrated information processing system which facilitates convenient handling of data. It allows for the automatic annotation of data produced from various published classification systems, with selected biological information derived from a database, for manual searching and browsing through these annotations, and for automatic generation of meaningful generalisations according to statistical criteria [117]. Furthermore, this particular system allows for the graphic presentation of ontologies and the display of hierarchical clustering results. TreeDyn is a tool based on data visualisation methods and dynamic graphics for the annotation of multiple phylogenetic or classification trees.

4 Bioinformatics and interactome

4.1 What is the interactome?

In the last few years, the deciphering of gene/protein function at a large scale to allow for a better understanding of cell functioning and organism development has stimulated the design of new analytical approaches, following both advances in methodology and current thinking. In this respect, bioinformatics methods have evolved 'in tune' with the way biologists have perceived gene/protein function [118, 119]. This is exemplified by the fact that the development of new computational methods, by allowing the decoding of the cellular, physiological and developmental function of gene/proteins at a large scale, has not only widened the field of investigation, but also more importantly, has brought about a novel, comprehensive and integrated understanding of gene/protein function and their interactions. This integrated view is nowadays more than ever an important part of the holistic concept of cell and organism functioning as currently portrayed by 'systems biology' (http://en.wikipedia.org/wiki/Systems_biology). Interestingly here also, the need for new bioinformatics tools is coming into prominence for identifying and analysing the phenomena and associated properties emerging from complex biological systems.

The last few years have witnessed the birth of new biological entities named interactomes. They correspond in an 'ideal world' to the complete set of protein–protein interactions existing between all the proteins of an organism. In reality they are far from complete since an unknown number of interactions are yet to be discovered. Current interactomes are only a part of the whole set of possible interactions occurring within an organism or between organisms. They are generally assembled from: (i) the results of large-scale two hybrid screens (LS-Y2H) (around 6000, 4000, 23 000 and 5500 interactions for yeast [120, 121], the nematode, *Caenorhabditis elegans* (Maupas) [122], *Drosophila* spp. [123, 124] and humans [125, 126], respectively); and (ii) the interactions identified by low-scale experiments described in the literature that may be eventually compiled in specialised databases (*e.g.* INTACT [127], MINT [128], HPRD [129], BIND [130]). Consequently, they do not reflect temporal influences because interactions are gathered from different cell types, tissues, development stages and types of experiment.

4.2 Bioinformatics methods developed in order to functionally investigate the interactome

Interactomes form large intricate networks leading to a renewed vision of cell biology as an integrated system. However, extracting and revealing the functional information

© 2006 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

they contain depends on our ability to analyse them in detail. For this, bioinformatics methods which partition the interaction network into functional modules have been proposed. These modules usually correspond to group of proteins involved in the same pathway, the same protein complex or the same cellular process.

Since interaction networks are represented by complex graphs in which nodes correspond to proteins and edges to the interactions, a number of these network analysis methods have been grounded on principles that derive from graph theory. Noticeably, a functional module or a class of protein that is functionally related and based on network analysis can be deduced from: (i) a search for graph regions particularly densely populated by interactions [131, 132]; (ii) the similarity between the shortest paths in the graph [133]; (iii) the progressive disconnection of the graph using a calculation of edge 'betweenness' [134, 135]; and (iv) the sharing of interactors [100, 136] or a combination thereof [102]. Some of these methods use the functional annotations of the protein (such as Gene Ontology annotations) to annotate the functional modules they predict. Based on the characteristics of a protein of unknown function to some of these annotated modules or classes, a putative function for such proteins can be proposed (Fig. 4 and [100]).

Currently we are in the period in which specialised methods are being developed to investigate these new biological entities, the so-called interactomes. But it is at the same time clear that the field starts to move forward: thus some of the previously cited methods have been implemented as softwares, 'plugins' or servers for a free use as bioinformatics tools by the 2-DE research community (for instance, MCODE in Cytoscape [137]; Prodistin (see Fig. 4) [138]) and start to be available as real bioinformatics tools as a direct result of the work of users. Although beyond the remit of this review, it is to be noted that the graphical representation of interactomes as very large graphs is also a real bioinformatics challenge successfully tackled by packages such as BioLayout [139], Cytoscape [137], Osprey [140] and Visant [141].

4.3 New prospects in interactome studies

4.3.1 Data integration

Interactomes are static structures which do not reflect temporal changes. Recently, the combination of interaction data with other data such as coexpression and phenotypic profiles allow the introduction of a dynamic aspect to the study of interactomes. From these new data, integration approaches have deepened our understanding of interactome structure by, for instance, (i) showing the existence of two types of highly connected proteins in interactomes, with respect to mRNA expressions [142]; and (ii) predicting 'molecular machines' involved in the embryogenesis of the nematode, *C. elegans* [143].

When evoking data integration, most of the works found in the literature use correlation coefficients such as Pearson or Bayesian network models. We suggest that graph theory-based methods for analysing interactomes can also be adapted to this problem by considering for example, graphs with weighted edges. In such graphs, an edge would be sustained by the existence of a detected physical interaction and the weight of the edge could reflect any type of shared feature, *e.g.* coexpression, colocalisation and coannotation. Interestingly, visualisation tools are also progressing in this direction by providing correlation interfaces [144]. Therefore, if one wishes to functionally analyse such weighted graphs, there is a need to develop a new graph theory-based method taking into account the novel dimension given to the interactome by the integration of functional data.

Like mRNA expression data or localisation data can be combined to protein–protein interactions to gain in biological relevance, the same goal can be achieved by using the results of interactome analysis in a reverse proteomics approach to interpret raw MS data. This was recently illustrated by Hinsby *et al.* [145] in an effort to assign a function to uncharacterised human nucleolar proteins. First, these authors have investigated the nucleolar interactome using one of the devoted tool previously cited [137]. This way, they identified clusters composed of bona fide nucleolar proteins mixed with others that have not been described previously as nucleolar. Then, second, they revisit the results of a large-scale nucleolus proteome MS study, by performing a targeted search for these putative novel nucleolar proteins in the MS data. Indeed, they were able to verify the presence of 11 of these proteins that were originally discarded due to their low score in the conservative unbiased MS search. Therefore, interactome and proteome analysis can be complementary.

4.3.2 Host–pathogen interactomes

Although the deciphering of the interactomes of the main model organisms is not yet complete, studies of the interactomes of pathogens are increasing. The first pathogens to be investigated in the past 5 years or so in terms of their interactomes were the hepatitis C virus [146] and the bacterium, *Helicobacter pylori* [147]. More recently still, the interactomes of the herpes viruses [148] and the malaria parasite, *Plasmodium falciparum*, [149] have been determined. This makes one believe that in the near future, as initiated by Uetz *et al.* [148], the docking of the interactomes of pathogens onto those of their hosts will be possible. The analysis of 'docked interactomes' is certainly a very promising and exciting aspect of interactomics because of its obvious potential impacts on human and animal health. But the fundamental questions that the docked interactomes allow us to ask are at least as exciting for people in other related fields, including those working on host–parasite interactions of metazoans . . . and as such, are urging upon us the need for bioinformatic tools to investigate interactomes *per se*, let alone docked ones!

© 2006 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

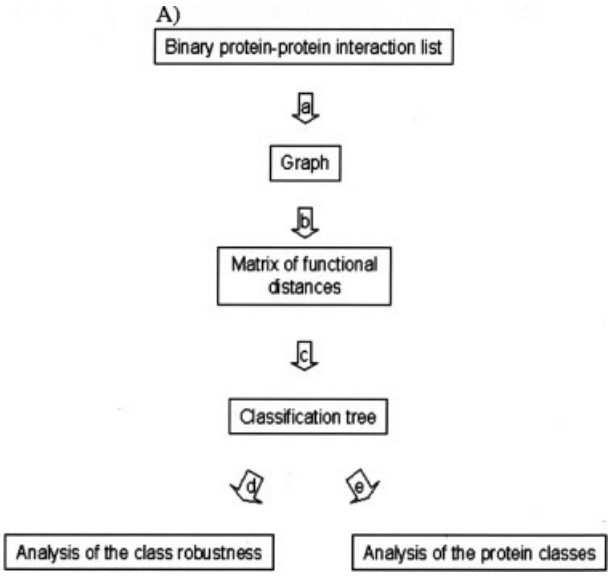
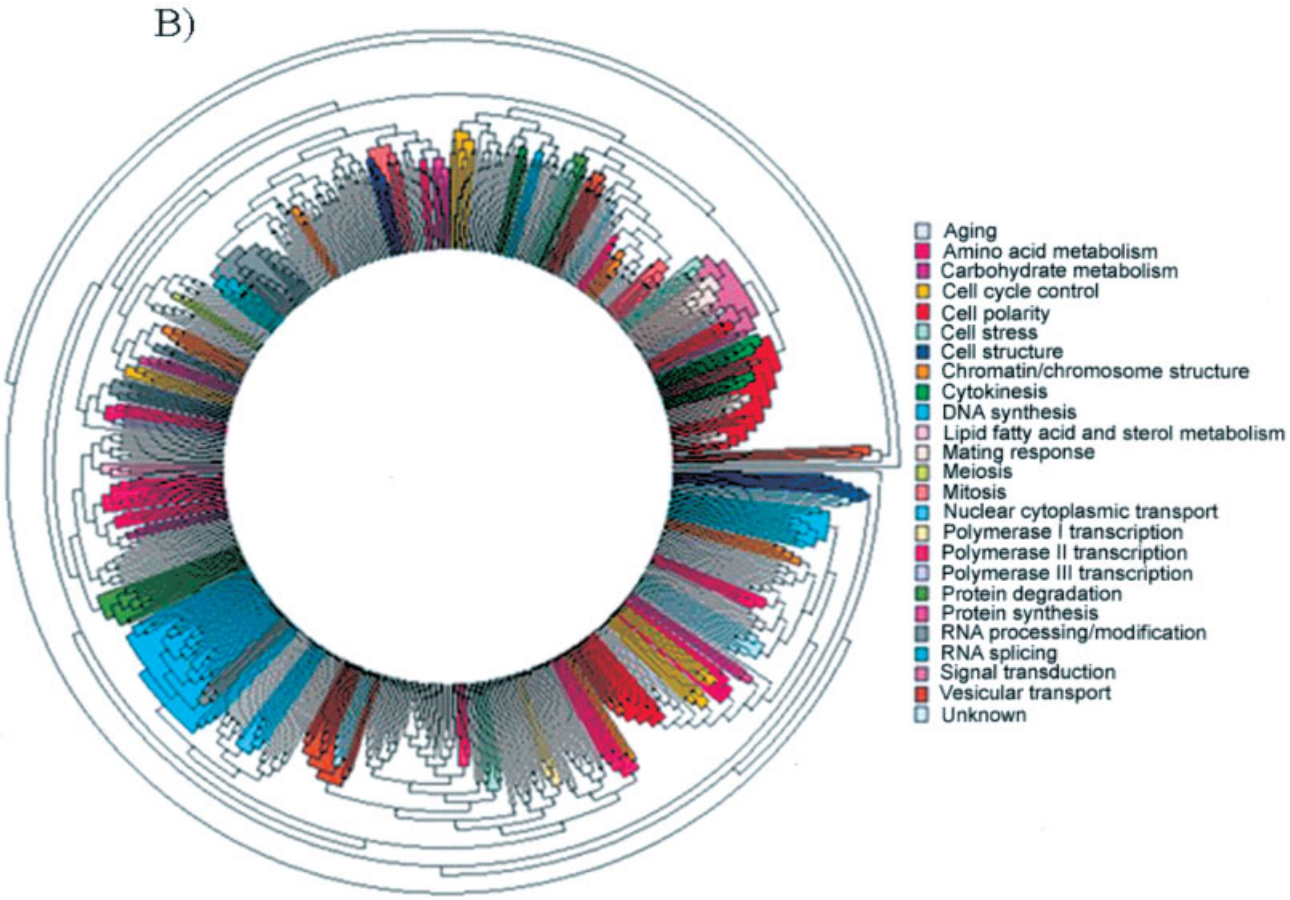


Figure 4. (A) Flowchart of PRODISTIN. (a) A graph is constructed from a list of binary protein–protein interactions. (b) A functional distance based on the identity of the shared interactors is calculated among all proteins. (c) The obtained distance matrix is used to build a classification tree, on which functional classes are subsequently determined and analysed by evaluating (d) their statistical robustness and (e) their biological relevance. (B) A functional classification tree for 602 yeast proteins computed with the PRODISTIN method. PRODISTIN classes have been coloured according to their corresponding 'Cellular Role', on the circular classification tree. Protein names have been omitted for the sake of clarity. We have showed that the clustering of the proteins reveals the biological process in which they are involved (for further details, see [100]).



5 Conclusions

Since the 1990s, 2-DE and MS have been successfully employed in a large number of studies to find and identify proteins involved in biological phenomena, *e.g.* immunity, response to environmental stresses, host–parasite interactions, *etc.* Even so, many studies have, as outlined above, revealed pitfalls in the approaches used. Probably, over the last two decades, mismatches of proteins were performed in a number of published proteomics studies. Thus, whatever

© 2006 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

the new technological advancements, more especially in 2-DE and in protein identification, it is apparent that proteomics researchers should attempt to improve their experimental design as well as take greater cognisance of, and have greater respect for, statistical approaches. This new attitude will surely improve the reliability of the data deriving from proteomics studies and will open the way for an enhanced comprehension of many biological mechanisms. In the near future, a greater amount of proteomics data will be available for many organisms and will in turn open up new prospects for interactome studies. By example, recently, the combination of the proteomics and interactome data on the human nuclear proteome permitted to assign function to 49 previously uncharacterised human nucleolar proteins and to reveal the first draft of the human ribosome biogenesis pathway [145]. In relation to 2-DE experiments, a promising research field is the study of the 'interactome' of organisms along with the instantaneous and the temporal interactomes resulting from the interaction of the proteomes between organisms, more especially as far as we are concerned, host–parasite interactions. Many new bioinformatics tools should be developed as a result of these new prospects in interactome studies. If so, the future in this area indeed appears to be a bright one, with the possibility that many complex protein–protein interactions that currently remain intractable will ultimately prove resolvable at several levels.

We are most grateful to the two anonymous referees for their valuable comments on the manuscript. We apologise to authors whose work was not cited owing to restrictions of space allowed. We thank Denis Sereno and Christophe Brugidou from IRD, Montpellier for their helpful comments on the manuscript. C. Lebarbenchon is supported by a grant from the Languedoc-Roussillon region and the Station Biologique de la Tour du Valat, whilst this work was supported by an ACI 'jeunes chercheurs' grant to F. Thomas.

6 References

[1] McKusick, V. A., *Genomics* 1997, *45*, 244–249.

[2] Snape, J. R., Maund, S. J., Pickford, D. B., Hutchinson, T. H., *Aquat. Toxicol.* 2004, *67*, 143–154.

[3] Coppel, R. L., Black, C. G., *Int. J. Parasitol.* 2005, *35*, 465–479.

[4] Hochstrasser, D. F., *Clin. Chem. Lab. Med.* 1998, *36*, 825–836.

[5] Vihinen, M., *Biomol. Eng.* 2001, *18*, 241–248.

[6] Lester, P. L., Hubbard, S. J., *Proteomics* 2002, *2*, 1392–1405.

[7] Boguski, M. S., McIntosh, M., *Nature* 2003, *422*, 233–237.

[8] Wheelock, A. M., Goto, S., *Expert. Rev. Proteomics* 2006, *3*, 129–142.

[9] Molloy, M. P., Brzezinski, E. E., Hang, J., McDowell, M. T. *et al.*, *Proteomics* 2003, *3*, 1912–1919.

[10] Carr, S., Aebersold, R., Baldwin, M., Burlingame, A. *et al.*, *Mol. Cell. Proteomics* 2004, *3*, 531–533.

[11] Karp, N. A., Kreil, D. P., Lilley, K. S., *Proteomics* 2004, *4*, 1421–1432.

[12] Hunt, S. M., Thomas, M. R., Sebastian, L. T., Pedersen, S. K. *et al.*, *J. Proteome Res.* 2005, *4*, 809–819.

[13] Karp, N. A., Lilley, K. S., *Proteomics* 2005, *5*, 3105–3115.

[14] Wilkins, M. R., Appel, R. D., Van Eyk, J. E., Chung, M. C. M. *et al.*, *Proteomics* 2006, *6*, 4–8.

[15] O'Farrell, P. H., *J. Biol. Chem.* 1975, *250*, 4007–4021.

[16] Hebert, B., *Electrophoresis* 1999, *20*, 660–663.

[17] Molloy, M. P., *Anal. Biochem.* 2000, *280*, 1–10.

[18] Patton, W. F., *Electrophoresis* 2000, *21*, 1123–1144.

[19] Mackintosh, J. A., Choi, H. Y, Bae, S. H., Veal, D. A. *et al.*, *Proteomics* 2003, *3*, 2273–2288.

[20] Simpson, R. J., *Proteins and Proteomics: A Laboratory Manual*, Cold Spring Harbor Laboratory Press, New York 2003.

[21] Van den Bergh, G., Arckens, L., *Curr. Opin. Biotech.* 2004, *15*, 38–43.

[22] Chevalier, F., Rodifal, V., Vanova, P., Bergoin, A. *et al.*, *Phytochemistry* 2004, *65*, 1499–1506.

[23] Zeng, L. W., Singh, R. S., *Genetics* 1993, *135*, 135–147.

[24] Sakoyama, Y., Okubo, S., *Dev. Biol.* 1981, *81*, 361–365.

[25] Klerk, H., Van Loon, L. C., *J. Exp. Biol.* 1991, *42*, 1295–1304.

[26] Cheney, C. M., Miller, K. G., Lang, T. J., Shearn, A., *Proc. Natl. Acad. Sci. USA* 1984, *81*, 6422–6426.

[27] Lee, C. Y., Charles, D., Bronson, D., Griffin, M. *et al.*, *Mol. Gen. Genet.* 1979, *176*, 303–311.

[28] Ohnishi, S., Leigh Brown, A. J., Voelker, R. A., Langley, C. H., *Genetics* 1981, *100*, 127–136.

[29] Kim, B. K., *Korean J. Genet.* 1988, *10*, 77–84.

[30] Goldman, D., Giri, P. R., O'Brien, S. J., *Proc. Natl. Acad. Sci. USA* 1987, *84*, 3307–3311.

[31] Spicer, G. S., *J. Mol. Evol.* 1988, *27*, 250–260.

[32] Choudhary, M., Coulthart, M. B., Singh, R. S., *Genetics* 1992, *130*, 843–853.

[33] Tastet, C., Bossis, M., Gauthier, J. P., Renault, L. *et al.*, *Nematology* 1999, *1*, 301–314.

[34] Moreno Da Cunha, M. J., Bossis, M., De Oliveira Abrantes, I. M., Newton, M. S. *et al.*, *Nematology* 2000, *2*, 461–471.

[35] Appel, R., Hochstrasser, D., Roch, C., Funk, M. *et al.*, *Electrophoresis* 1988, *9*, 136–142.

[36] Pun, T., Hochstrasser, D. F., Appel, R. D., Funk, M. *et al.*, *Appl. Theor. Electrophor.* 1988, *1*, 3–9.

[37] Lemkin, P. F., *Electrophoresis* 1997, *18*, 461–470.

[38] Rabilloud, T., *Proteomics* 2002, *2*, 3–10.

[39] Anderson, N. L., Taylor, J., Scandora, A. E., Coulter, B. P. *et al.*, *Clin. Chem.* 1981, *27*, 1807–1820.

[40] Appel, R. D., Hochstrasser, D. F., Funk, M., Vargas, J. R. *et al.*, *Electrophoresis* 1991, *12*, 722–735.

[41] Marengo, E., Robotti, E., Gianotti, V., Righetti, P. G., *Ann. Chim.* 2003, *93*, 105–116.

[42] Marengo, E., Robotti, E., Righetti, P. G., Antonucci, F., *J. Chromatogr. A* 2003, *1004*, 13–28.

[43] Marengo, E., Robotti, E., Gianotti, V., Righetti, P. G. *et al.*, *Electrophoresis* 2003, *24*, 225–236.

© 2006 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

[44] Marengo, E., Robotti, E., Cecconi, D., Scarpa, A. *et al.*, *FUZ-Z-IEE 2004 Proceedings*, Budapest, Hungray 2004, *1*, 359–364.

[45] Barrett, J., Brophy, P. M., Hamilton, J. V., *Int. J. Parasitol.* 2005, *35*, 543–553.

[46] Marengo, E., Robotti, E., Antonucci, F., Cecconi, D. *et al.*, *Proteomics* 2005, *5*, 654–666.

[47] Raman, G., Cheung, A., Marten, M. R., *Electrophoresis* 2002, *23*, 2194–2202.

[48] Rosengren, A. T., Salmi, J. M., Aittokallio, T., Westerholm, J. *et al.*, *Proteomics* 2003, *3*, 1936–1946.

[49] Rogers, M., Graham, J., Tonge, R. P., *Proteomics* 2003, *3*, 879–886.

[50] Rogers, M., Graham, J., Tonge, R. P., *Proteomics* 2003, *3*, 887–896.

[51] Zhan, X., Desiderio, D. M., *Electrophoresis* 2003, *24*, 1834–1846.

[52] Beranova-Giorgianni, S., *Trends Anal. Chem.* 2003, *22*, 273–281.

[53] Scheler, C., Lamer, S., Pan, Z., Li, X. *et al.*, *Electrophoresis* 1998, *19*, 918–927.

[54] Hoving, S., Voshol, H., van Oostrum, J., *Electrophoresis* 2000, *22*, 2865–2871.

[55] Tonge, R., Shaw, J., Middleton, B., Rowlinson, R. *et al.*, *Proteomics* 2001, *1*, 377–396.

[56] Nishihara, J. C., Champion, K. M., *Electrophoresis* 2002, *23*, 2203–2215.

[57] Berggren, K. N., Schulenberg, B., Lopez, M. F., Steinberg, T. H. *et al.*, *Proteomics* 2002, *2*, 486–498.

[58] Ponton, F., Lebarbenchon, C., Lefèvre, T., Thomas, F. *et al.*, *Parasitology* 2006, DOI: 10.1017/50031182006000904

[59] Mahon, P., Dupree, P., *Electrophoresis* 2001, *22*, 2075–2805.

[60] Raman, B., Cheung, A., Marten, M. R., *Electrophoresis* 2002, *23*, 2194–2202.

[61] Krell, D. P., Karp, N. A., Lilley, K. S., *Bioinformatics* 2004, *20*, 2026–2034.

[62] Karp, N. A., Griffin, J. L., Lilley, K. S., *Proteomics* 2005, *5*, 81–90.

[63] Scherrer, B., *Biostatistique*, Gaëtan Morin Editeur, Boucherville, Canada 1984.

[64] Tomassone, R., Dervin, C., Mason, J. P., *Biométrie: Modélisation de phénomènes biologiques*, Masson, Paris 1993, 544p.

[65] Fodor, I. K., Nelson, D. O., Alegria-Hartman, M., Robbins, K. *et al.*, *Bioinformatics* 2005, *21*, 3733–3740.

[66] Meunier, B., Bouley, J., Piec, I, Bernanrd, C. *et al.*, *Anal. Biochem.* 2005, *340*, 226–230.

[67] Shapiro, S. S., Wilk, M. B., *Biometrika* 1965, *52*, 591–611.

[68] Huber, W., von Heydebreck, A., Sultmann, H., Poustka, A. *et al.*, *Bioinformatics* 2002, *18*, S96–S104.

[69] Gustafsson, J. S., Ceasar, R., Glasbey, C. A., Blomberg, A. *et al.*, *Proteomics* 2004, *4*, 3791–799.

[70] Benjamini, Y., Hochberg, Y., *J. Royal Statist. Soc. B* 1995, *57*, 289–300.

[71] Barrett, J., Jefferies, J. R., Brophy, P. M., *Parasitol. Today* 2000, *9*, 400–403.

[72] Ashton, P. D., Curwen, R. S., Wilson, R. A., *Trends Parasitol.* 2001, *17*, 198–202.

[73] Biron, D. G., Moura, H., Marché, L., Hughes, A. L. *et al.*, *Trends Parasitol.* 2005, *21*, 162–168.

[74] Biron, D. G., Joly, C., Galeotti, N., Ponton, F. *et al.*, *Behav. Process* 2005, *68*, 249–253.

[75] Francis, F., Gerkens, P., Harmel, N., Mazzucchelli, G. *et al.*, *Insect Biochem. Mol.* 2006, *36*, 219–227.

[76] Engström, Y., Loseva, O., Theopold, U., *Trends Biotech.* 2004, *22*, 600–605.

[77] Hayman, M. W., Przyborski, S. A., *Biochem. Bioph. Res. Commun.* 2004, *316*, 918–923.

[78] Bandara, L. R., Kennedy, S., *Drug Discov. Today* 2002, *7*, 411–418.

[79] Wetmore, B. A., Merrick, B. A., *Toxicol. Pathol.* 2004, *32*, 619–642.

[80] Wilkins, M. R., Williams, K. K., *J. Theor. Biol.* 1997, *186*, 7–15.

[81] Mathesius, U., Imin, N., Chen, H., Djordjevic, M. A. *et al.*, *Proteomics* 2002, *2*, 1288–1303.

[82] Hanash, S., *Nature* 2003, *422*, 226–232.

[83] Karas, M., Hillenkamp, F., *Anal. Chem.* 1988, *60*, 2301–2303.

[84] Smith, R. D., Loo, J. A., Edmonds, C. G., Baringa, C. J. *et al.*, *Anal. Chem.* 1990, *62*, 882–889.

[85] Hillenkamp, F., Karas, L., Beavis, R. C., Chait, B. T., *Anal. Chem.*, 1991, *63*, 1193A–1203A.

[86] Pappin, D. J. C., Hojrup, P., Bleasby, A. J., *Curr. Biol.* 1993, *3*, 327–332.

[87] Roepstorff, P., *Curr. Opin. Biotechnol.* 1997, *8*, 6–13.

[88] Jennings, K. R., *Chromatogr. Separ. Technol.* 2000, *11*, 18–22.

[89] Aebersold, R., Mann, M., *Nature* 2003, *422*, 198–207.

[90] Sali, A., Glaeser, R., Earnest, T., Baumeister, W., *Nature* 2003, *422*, 216–225.

[91] Wilkins, M. R., Sanchez, J. C., Gooley, A. A., Appel, R. D. *et al.*, *Biotechnol. Genet. Eng. Rev.* 1996, *13*, 19–50.

[92] Mann, M., Wilm, M., *Anal. Chem.* 1994, *66*, 4390–4399.

[93] Wilkins, M. R., Gasteiger, E., Wheeter, C. H., Sanchez, I. *et al.*, *Electrophoresis* 1998, *19*, 3199–3206.

[94] Nesvizhskii, A. I., Aebersold, R., *DDT* 2004, *9*, 173–181.

[95] Edman, P., *Thromb. Diath. Haemorrhag.* 1964, *13*, 17–20.

[96] Edman, P., Begg, G., *Eur. J. Biochem.* 1967, *1*, 80–91.

[97] Henzel, W. J., Watanabe, C., Stults, J. T., *J. Am. Soc. Mass Spectrom.* 2003, *14*, 931–942.

[98] Carr, S., Aebersold, R., Baldwin, M., Burlingame, A. *et al.*, *Mol. Cell. Proteomics* 2004, *3*, 531–533.

[99] Wilkins, M. R., Gooley, A. A., in: Wilkins, M. R., Williams, K. L., Appel, R. D., Hochstrasser, D. F. (Eds.), *Protein Identification in Proteome Project*, Springer-Verlag, Berlin 1997, pp. 35–64.

[100] Brun, C., Martin, D., Chevenet, F., Wojcik, J. *et al.*, *Genome Biol.* 2003, *5*, R6.

[101] Tyers, M., Mann, M., *Nature* 2003, *422*, 193–197.

[102] Brun, C., Herrmann, C., Guenoche, A., *BMC Bioinformatics* 2004, *5*, 95.

[103] Gasteiger, E., Hoogland, C., Gattiker, A., Duvaud, S. *et al.*, in: Walker, J. M. (Ed.), *Protein Identification and Analysis Tools of the Expasy Server*, Human Press Inc., Totowa, NJ 2005, pp. 571–607.

© 2006 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

[104] Choudhary, J. S., Blackstock, W. P., Creasy, D. M., Cottrell, J. S., *Trends Biotechnol.* 2001, *19*, S17–S22.

[105] Keller, A., Nesvizhskii, A. I., Kolker, E., Aebersold, R., *Anal. Chem.* 2002, *74*, 5383–5392.

[106] Liska, A. J., Shevchenko, A., *Proteomics* 2003, *3*, 19–28.

[107] Lester, P. J., Hubbard, S. J., *Proteomics* 2002, *2*, 1392–1405.

[108] Habermann, B., Oegema, J., Sunyaev, S., Shevchenko, S., *Mol. Cell. Proteomics* 2004, *3*, 238–249.

[109] Pertsemlidis, A., Fondon, J. W. III, *Genome Biol.* 2001, *2*, 1–10.

[110] Genereux, D. P., Logsdon, J. M., Jr., *Trends Genet.* 2003, *19*, 191–195.

[111] Liska, A. J., Sunyaev, S., Shilov, I. N., Schaeffer, D. A. *et al.*, *Proteomics* 2005, *5*, 4118–4122.

[112] Neubauer, G., King, A., Rappsilber, J., Calvio, C. *et al.*, *Nature* 1998, *20*, 46–50.

[113] Tomarev, S. I., Zinovieva, R. D., *Nature* 1988, *336*, 86–88.

[114] Barrett, J., *Int. J. Biochem. Cell. Biol.* 2001, *33*, 105–117.

[115] Ostrowski, M., Fegatella, F., Wassinger, V., Guilhaus, M. *et al.*, *Proteomics* 2004, *4*, 1779–1788.

[116] Seo, J., Schneiderman, B., in: Bederson, B. B., Shneiderman, B. (Eds.), *The Craft of Information Visualization*, Morgan Kaufmann, San Francisco, USA, 2003, pp. 334–340.

[117] Pasquier, C., Girardot, F., Jevardat de Fombelle, K., Christen, R., *Bioinformatics* 2004, *20*, 2636–2643.

[118] Brun, C., Baudot, A., Guénoche, A., Jacq, B, in: Kamp, R. M., Calvete, J. J., Choli-Papadopoulou, T. (Eds.), *Principles and Practices – Methods in Proteome and Protein Analysis*, Springer-Verlag, Berlin, Heidelberg 2004, pp. 103–124.

[119] Brun, C., Baudot, A., Jacq, B, in: Dunn, M., Jorde, L., Little, P., Subramaniam, S. (Eds.), *Encyclopedia of Genomics, Proteomics and Bioinformatic*, John Wiley & Sons Limited, Weinheim 2005, Vol. 5, Chapter 46, pp. 1–7.

[120] Ito, T., Chiba, T., Ozawa, R., Yoshida, M. *et al.*, *Proc. Natl. Acad. Sci. USA* 2001, *98*, 4569–4574.

[121] Uetz, P., Giot, L., Cagney, G., Mansfield, T. A. *et al.*, *Nature* 2000, *403*, 623–627.

[122] Li, S., Armstrong, C. M., Bertin, N., Ge, H. *et al.*, *Science* 2004, *303*, 540–543.

[123] Formstecher, E., Aresta, S., Collura, V., Hamburger, A. *et al.*, *Genome Res.* 2005, *15*, 376–384.

[124] Giot, L., Bader, J. S., Brouwer, C., Chaudhuri, A. *et al.*, *Science* 2003, *302*, 1727–1736.

[125] Rual, J. F., Venkatesan, K., Hao, T., Hirozane-Kishikawa, T. *et al.*, *Nature* 2005, *437*, 1173–1178.

[126] Stelzl, U., Worm, U., Lalowski, M., Haenig, C. *et al.*, *Cell* 2005, *122*, 957–968.

[127] Hermjakob, H., Montecchi-Palazzi, L., Lewington, C., Mudali, S. *et al.*, *Nucleic Acids Res.* 2004, *32*, D452–D455.

[128] Zanzoni, A., Montecchi-Palazzi, L., Quondam, M., Ausiello, G. *et al.*, *FEBS Lett.*, 2002, *513*, 135–140.

[129] Peri, S., Navarro, J. D., Amanchy, R., Kristiansen, T. Z. *et al.*, *Genome Res.* 2003, *13*, 2363–2371.

[130] Alfarano, C., Andrade, C. E., Anthony, K., Bahroos, N. *et al.*, *Nucleic Acids Res.* 2005, *33*, D418–D424.

[131] Bader, G. D., Hogue, C. W., *BMC Bioinformatics*, 2003, *4*, 2.

[132] Spirin, V., Mirny, L. A., *Proc. Natl. Acad. Sci. USA* 2003, *100*, 12123–12128.

[133] Rives, A. W., Galitski, T., *Proc. Natl. Acad. Sci. USA* 2003, *100*, 1128–1133.

[134] Girvan, M., Newman, M. E., *Proc. Natl. Acad. Sci. USA* 2002, *99*, 7821–7826.

[135] Dunn, R., Dudbridge, F., Sanderson, C. M., *BMC Bioinformatics* 2005, *6*, 39.

[136] Samanta, M. P., Liang, S., *Proc. Natl. Acad. Sci. USA* 2003, *100*, 12579–12583.

[137] Shannon, P., Markiel, A., Ozier, O., Baliga, N. S. *et al.*, *Genome Res.* 2003, *13*, 2498–504.

[138] Baudot, A., Martin, D., Mouren, P., Chevenet, F. *et al.*, *Bioinformatics* 2006, *22*, 248–250.

[139] Goldovsky, L., Cases, I., Enright, A. J., Ouzounis, C. A., *Appl. Bioinformatics* 2005, *4*, 71–74.

[140] Breitkreutz, B. J., Stark, C., Tyers, M., *Genome Biol.* 2003, *4*, R22.

[141] Hu, Z., Mellor, J., Wu, J., Yamada, T. *et al.*, *Nucleic Acids Res.* 2005, *1*, 33.

[142] Han, J. D., Bertin, N., Hao, T., Goldberg, D. S. *et al.*, *Nature* 2004, *1*, 43088–43093.

[143] Gunsalus, K. C., Ge, H., Schetter, A. J., Goldberg, D. S. *et al.*, *Nature* 2005, *436*, 861–865.

[144] Tarassov, K., Michnick, S. W., *Genome Biol.* 2005, *6*, R115.

[145] Hinsby, A. M., Kiemer, L., Karlberg, E. O., Lage, K. *et al.*, *Mol. Cell* 2006, *22*, 285–295.

[146] Flajolet, M., Rotondo, G., Daviet, L., Bergametti, F. *et al.*, *Gene* 2000, *242*, 369–379.

[147] Rain, J. C., Selig, L., De Reuse, H., Battaglia, V. *et al.*, *Nature* 2001, *409*, 211–215.

[148] Uetz, P., Dong, Y. A., Zeretzke, C., Atzler, C. *et al.*, *Science* 2006, *311*, 239–242.

[149] LaCount, D. J., Vignali, M., Chettier, R., Phansalkar, A. *et al.*, *Nature* 2005, *438*, 103–107.

© 2006 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim