Journal of Experimental Botany, Vol. 56, No. 410,
Making Sense of the Metabolome Special Issue, pp. 273–286, January 2005
doi:10.1093/jxb/eri068 Advance Access publication 23 December, 2004

Journal of Experimental Botany

Metabolite profiling of fungi and yeast: from phenotype to metabolome by MS and informatics

Jørn Smedsgaard* and Jens Nielsen

Center for Microbial Biotechnology, BioCentrum-DTU, Technical University of Denmark, Søltofs Plads Building 221, DK-2800 Kgs Lyngby, Denmark

Received 20 May 2004; Accepted 28 October 2004

Abstract

Filamentous fungi and yeast from the genera *Saccharomyces*, *Penicillium*, *Aspergillus*, and *Fusarium* are well known for their impact on our life as pathogens, involved in food spoilage by degradation or toxin contamination, and also for their wide use in biotechnology for the production of beverages, chemicals, pharmaceuticals, and enzymes. The genomes of these eukaryotic micro-organisms range from about 6000 genes in yeasts (*S. cerevisiae*) to more than 10 000 genes in filamentous fungi (*Aspergillus* sp.). Yeast and filamentous fungi are expected to share much of their primary metabolism; therefore much understanding of the central metabolism and regulation in less-studied filamentous fungi can be learned from comparative metabolite profiling and metabolomics of yeast and filamentous fungi. Filamentous fungi also have a very active and diverse secondary metabolism in which many of the additional genes present in fungi, compared with yeast, are likely to be involved. Although the 'blueprint' of a given organism is represented by the genome, its behaviour is expressed as its phenotype, i.e. growth characteristics, cell differentiation, response to the environment, the production of secondary metabolites and enzymes. Therefore the profile of (secondary) metabolites–fungal chemodiversity–is important for functional genomics and in the search for new compounds that may serve as biotechnology products. Fungal chemodiversity is, however, equally efficient for identification and classification of fungi, and hence a powerful tool in fungal taxonomy. In this paper, the use of metabolite profiling is discussed for the identification and classification of yeasts and filamentous fungi, functional analysis or discovery by integration of high performance analytical methodology, efficient data handling techniques and core concepts of species, and intelligent screening. One very efficient approach is direct infusion Mass Spectrometry (diMS) integrated with automated data handling, but a full metabolic picture requires the combination of several different analytical techniques.

Key words: Fungi, metabolic engineering, screening strategy, species model.

Introduction

Filamentous fungi and yeasts (Ascomycetes belonging to the kingdom Mycota) are important micro-organisms in the environment as they play an essential role in connection with plant growth and are of importance in carbon recycling in nature. Yeasts that also belong to the fungal kingdom have been used for fermentation of food and beverages since ancient times and are today widely used for industrial production of chemicals, pharmaceuticals, and proteins. Filamentous fungi are also used extensively in biotechnology as they can produce a wide range of chemicals that are used as food ingredients, pharmaceuticals, enzymes, and solvents. Besides the beneficial use of both yeast and filamentous fungi in biotechnology, these micro-organisms are also involved in food spoilage and many species are also pathogens to plants, animals, and humans (Smith and Solomons, 1994).

Compared with plants, filamentous fungi and yeasts show a lower degree of cellular differentiation, but still they express a complex metabolism resulting in the production of a broad range of metabolites (secondary metabolites) (Turner, 1971; Cole and Cox, 1981; Turner and Aldridge, 1983; Nielsen and Smedsgaard, 2003; Cole and Schweikert, 2003*a*, *b*; Cole *et al.*, 2003; Frisvad *et al.*, 2004) and extracellular enzymes. This very high metabolic diversity

* To whom correspondence should be addressed. Fax: +45 4588 4922. E-mail: js@biocentrum.dtu.dk

has been actively exploited for many years and many metabolites produced by filamentous fungi are bioactive compounds, and have found use as antibiotics, cholesterol-lowering agents, antitumour agents, and immunosuppresors (Newman *et al.*, 2003). In terms of biotechnological application filamentous fungi and yeast have the advantage of being relatively easy to grow in fermenters and they are therefore well-suited for large-scale industrial production.

Discovery, understanding and utilization of fungi require a combination of knowledge and techniques from several scientific disciplines including: taxonomy, analytical chemistry, ecology, biology, genetic engineering, fermentation, and informatics. Efficient use of the information generated requires a working hypothesis, together with a strategy to utilize the information.

Metabolite profiling (or metabolome analysis) represents a tool that finds common application in all aspects of discovery, understanding and utilization, and hence it represents a focal point in studies of fungal taxonomy and physiology. The analytical methodologies for metabolite profiling have been extensively discussed in the literature and reviewed in several papers and books (Pramanik *et al.*, 2002; Harrigan and Goodacre, 2003; Villas-Bôas *et al.*, 2004). In this paper the focus will be on introducing a practical working concept of species which will help to facilitate an understanding of species, phenotype (in the case of fungi through expression by production of secondary metabolites) and functional genomics, particularly with the objective of discovering bio- and chemical diversity from analytical data. The species concept can, besides its obvious use in taxonomy, also form the basis for intelligent screening strategies to exploit the biotechnological potentials of fungi for the production of specific chemical structures or classes of chemicals. In this paper this strategy will be illustrated using the results from studies of *Penicillium* species. Furthermore, the use of metabolite profiling for designing novel strategies to develop more efficient cell factories through metabolic engineering, illustrated with data from yeast. Throughout the discussion attention will be drawn to some of the analytical and informatics' problems associated with the application of metabolic profiling.

The species concept

The simplified and practical 'concept of species' as illustrated in Fig. 1A is based on the core dogma that species exist and can be delimited and described through a *combined* use of the outward directed features of the organism (phenotypic characters) *together* with its relation to the environment (ecology) *and* its inheritance (phylogenetics or genome). This conceptual model serves several goals: to delimit and identify species efficiently (classification and taxonomy) (Fig. 1A), to discover new species (biodiversity) (Fig. 1B), to discover useful features across species, and to discover products or pathways of biotechnological interest (Fig. 1C). Furthermore, the model illustrates the working space for genetic engineering. Although the species concept is a matter of much debate which is outside the scope of this paper, the concept of species presented here includes several of the elements from the current debate (for example, the comments by Blaxter and Floyd, 2003; Dunn, 2003; Agosti, 2003) except for *time or evolution* which is not as relevant in practical biotechnology.

Historically, classification has been based on the outwardly directed features found on the first two axes: the phenetics, which includes all the features that can be seen or measured directly on the organism, for example, metabolite and enzyme production and on the second axis describing the ecological behaviour, for example, growth characteristics. This is illustrated in the polyphasic taxonomy where many characters are used in combination by, for example, Frisvad (1998) and Frisvad and Samson (2004) for the species within the *Penicillium* subgenus *Penicillium*. These two axes still represent the basis of fungal taxonomy. The

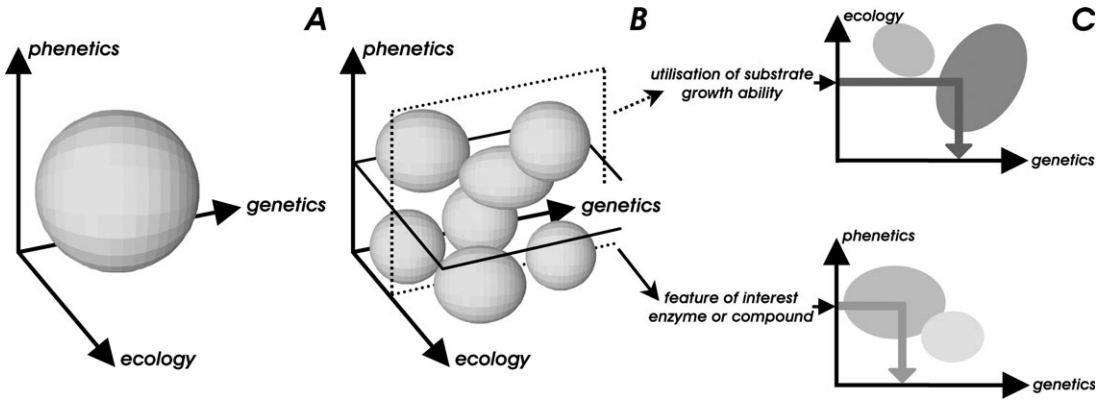

Fig. 1. The concept of species delimited by phenotypic characters, ecological behaviour, and genome is illustrated to the left (A). Looking at a population of species, where the number of these and their distribution represents the biodiversity, this population can be searched for features with biotechnological relevance, for example, metabolite production or substrate utilization (B, C). There is of course a linkage between phenetics and ecology (survival) which is of importance for ecological and biodiversity studies.

third axis delimited species by their genome, either in part or using the whole genome, for example, as discussed by Tautz *et al*. (2003) and reviewed by Cuarro *et al*. (1999) and also by Taylor *et al*. (2000). The genome axis reflects the heritage (phylogenetics) of a species, but it contains limited information about current ecological behaviour and other outwardly directed features (e.g. its morphology).

All three axes of the species concept (phenetics, ecology, and genetics) are strongly linked and somewhat redundant, but they are equally important in the context of describing, understanding, and classifying species. Systematics based on phenetics, i.e. observation of many different outwardly directed features works well for some organisms like filamentous fungi (Frisvad, 1998), whereas ecological studies involving the testing of growth patterns and substrate utilization are more efficient in yeast systematics. Genetics (phylogenetics) have the potential to give a clear classification of all organisms, however, as large parts of the genome can be quite similar within some genera it is necessary that the right genes are chosen carefully or that a sufficiently large part of the genome is sequenced with high resolution to give a usable delimitation of different fungal species (Samson *et al.*, 2004).

The conceptual species model serves several purposes, such as to illustrate the linkage of phenotypic characterization to functional genomics, for example, assigning function to genes based on detailed phenotypic analysis like metabolite profiling (Fig. 1B) or to assign ecological information to the genome (Fig. 1C). Thus, by screening a population for specific features, for example, the production of a metabolite or enzyme or those that can use a specific substrate or growth in a specific environment, it is possible to search for the corresponding genes in the relevant species. As the feature space spanned by phenetics and ecology is larger than that normally considered in traditional taxonomy a feature search will be much aided if the classification of species is based on a comprehensive and clear characterization that also includes metabolite profiling. Finally, the conceptual species model is valuable in the emerging field of functional biodiversity, where the function and genomic origin of features (e.g. the production of a metabolite) found in one organism can be explained by the occurrence in closely related organisms.

The chemical network

The integrated biochemistry of a cell can be viewed through the linkage of the different omes' as illustrated in Fig. 2. Linking gene function to a specific metabolic capabilities or metabolite production is crucial for understanding and development of biotechnological processes. Transcriptional profiling and proteomics are becoming routine techniques in many laboratories, and functions are assigned to genes either as a result of direct proof by comparative annotation. However, it is still not always clear whether a given gene

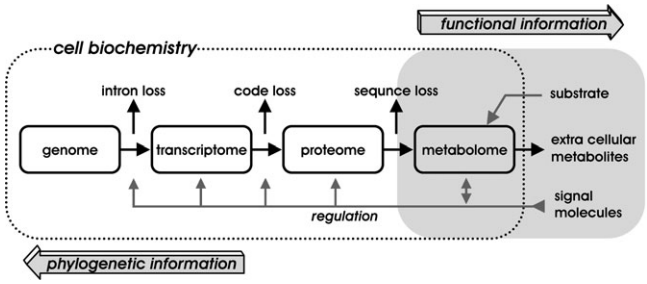

Fig. 2. Cell biochemistry (or biological informatics) is another way to illustrate the basic linkage of the inherited phylogenetic information (genes) to the observed biochemical phenotype metabolites, which form the basis of metabolic engineering.

may exert several functions or if it is inactive. It is therefore quite often necessary to perform a detailed phenotypic characterization in different growth conditions.

Metabolomics represents a newer complementary technique to functional genomics as it provides integrative information, i.e. a large number of genes may be involved in the production of one metabolite. Several different definitions have been proposed on metabolomics (Fiehn, 2002; Sumner *et al.*, 2003), but a practical definition is: 'The complete pool of small metabolites in a cell at any given time.' It is important to realize that there is not always (in fact rarely for the secondary metabolites) a one-to-one relationship between a gene and a metabolite, and the metabolite levels are therefore usually a complex result of the expression of many genes and the function of many enzymes. It is therefore inherently difficult to interpret the patterns of the metabolites, and particularly to infer something about gene functions based on metabolite profiling. Obviously, metabolomics study relies heavily on advanced analytical techniques to determine the many metabolites in one sample; sometimes it is also necessary to quantify the metabolite levels (Fiehn, 2002; Sumner *et al.*, 2003; Nielsen *et al.*, 2004; Kell, 2004). Profiling of secondary metabolites is also an important tool for the classification of filamentous fungi, as illustrated by Frisvad and co-workers over many years (see Frisvad, 1998, and the references in the practical example below).

Metabolic engineering

When desired features or organisms have been identified, it is often desirable to manipulate and/or transfer these features to other production hosts that are more suited for industrial production or change the organism to one better suited for the production environment and increased yield. A comprehensive introduction to this field can be found in Stephanopoulos *et al.* (1998). An example of a widely used cell factory is the yeast *Saccharomyces cerevisiae*, for which the complete genome is known and many detailed functional genomics studies have been carried out. Based on genomic information the metabolic network of this yeast

has been reconstructed (Föster *et al.*, 2003; Famili *et al.*, 2003). An attractive feature of this organism is that it is easy to perform genetic modification of this yeast, and it is therefore possible to engineer the metabolism and thereby exploit the organism as a host for the industrial production of many different chemicals. Filamentous fungi have larger genomes and represent rich sources for a variety of natural products; but unfortunately the pathways for most of these have not yet been elucidated. So far, the genome has only been sequenced for a few species of filamentous fungi (Hofmann *et al.*, 2003), and there has not been a detailed metabolic reconstruction of any filamentous fungi. However, based on physiological information and diverse sequence information, the central carbon metabolism of *Aspergillus niger*, which is extensively used for industrial production, has been reconstructed (David *et al.*, 2003). These reconstructed metabolic maps represent valuable information about the metabolome in these organisms. Together with the discovery of chemical diversity and useful functionality, the metabolic maps allow the design of novel cell factories for the production of different chemicals, for example, the production of polyketides by yeast.

Practical examples

Diversity and discovery in Penicillium *by profiling and informatics*

Filamentous fungi are able to express a fascinating chemical diversity through secreted metabolites. It has been estimated that more than 10 000 secondary metabolites may be present in *Aspergillus* and *Penicillium* (Jens C Frisvad, personal communication) where perhaps less than 10% of these metabolites are known. Several extensive reviews aim to compile knowledge about metabolite production by fungi (Turner, 1971; Cole and Cox, 1981; Turner and Aldridge, 1983; Cole and Schweikert, 2003*a*, *b*; Cole *et al.*, 2003). Taxonomy has always been considered difficult within these genera and something that can only be done by experts. There are, therefore, numerous misidentifications of species in the literature. This has led to erroneous postulates about metabolite production and also to the rejection of the usability of metabolites in classification. If the genus *Penicillium* is studied, the profiles of secondary metabolites have, for more than 20 years, proved to be a very efficient tool in classification and taxonomy (Frisvad and Filtenborg, 1989; Frisvad, 1998). The most complete work has been done for classification of the terverticillate penicillica (*Penicillium* subgenus *Penicillium*) where secondary metabolites have been used in several studies using different analytical approaches: TLC (Filtenborg *et al.*, 1983), HPLC (Svendsen and Frisvad, 1994; Nielsen *et al.*, 1999) and direct infusion ESI-MS (Smedsgaard and Frisvad, 1997; Smedsgaard *et al.*, 2004). This large diverse group of terverticillate penicillia which currently include 58 accepted species has recently been reviewed by Frisvad and Samson (2004) and is expected to include many not yet described species (Jens C Frisvad, personal communication). The general methodology for metabolite profiling of fungi has been reviewed recently by Nielsen *et al*. (2004) and metabolic reference information can be found in Nielsen and Smedsgaard (2003) and Frisvad *et al*. (2004).

Chemical image analysis

While the analytical protocols today are quite efficient the use of metabolite profiling relies on expert data evaluation which has been difficult to automate. First, consider the results from classical reversed phase HPLC analysis with diode array detection of a plug extract of a *Penicillium* species (Smedsgaard, 1997*a*) where full UV-spectra are collected at regular intervals as shown on Fig. 3.

Such analysis provides a wealth of information; in the retention time domain chromatograms can be extracted at different wavelengths, enhancing the specificity as shown by the traces in Fig. 3B and C and gaining information about compound polarity. At each retention time point the UV spectrum (Fig. 3D, E) can be extracted with the UV-spectral properties of the eluting component. Traditionally these data are evaluated by the detection of chromatographic peaks on the chromatograms (Fig. 3B, C) and their identity assigned by the combined use of the retention time and the UV-spectrum, together with reference information. This process can be semi-automated, but relies on peak detection, selection of relevant peaks, and libraries. However, the full HPLC data matrix can also be viewed as a 'chemical image' of the sample as shown in Fig. 3A where absorbance has been given a colour at each (retention time, wavelength) point. This topographic greyscale map combines and retains the information from the analysis without peak detection. Chromatographic images as shown in Fig. 3A can be compared automatically using techniques from image analysis. However, it is crucial to align the chromatograms before analysis to compensate for the minor shift in retention time from analysis to analysis. A very efficient way to do this is by the correlation optimized warping technique (COW), using all the spectral information (Nielsen *et al.*, 1998). The result of COW aligning is shown in Fig. 4 (just one chromatographic trace is shown from each sample) where two different isolates of the same species have been analysed.

Figure 4B shows a near perfect match of the chromatograms, compared with the raw chromatograms on Fig. 4A.

Analysing 45 isolates of eight closely related species by HPLC-UV and using chemical image analysis (CIA) (Nielsen *et al.*, 1998) on all 45 aligned data matrices, the overall similarity between these chromatograms can be calculated and this enabled classification of the different species by cluster analysis as shown in Fig. 5.

The dendrograms show a good classification of the eight species included in the study and only five of the 45 isolates

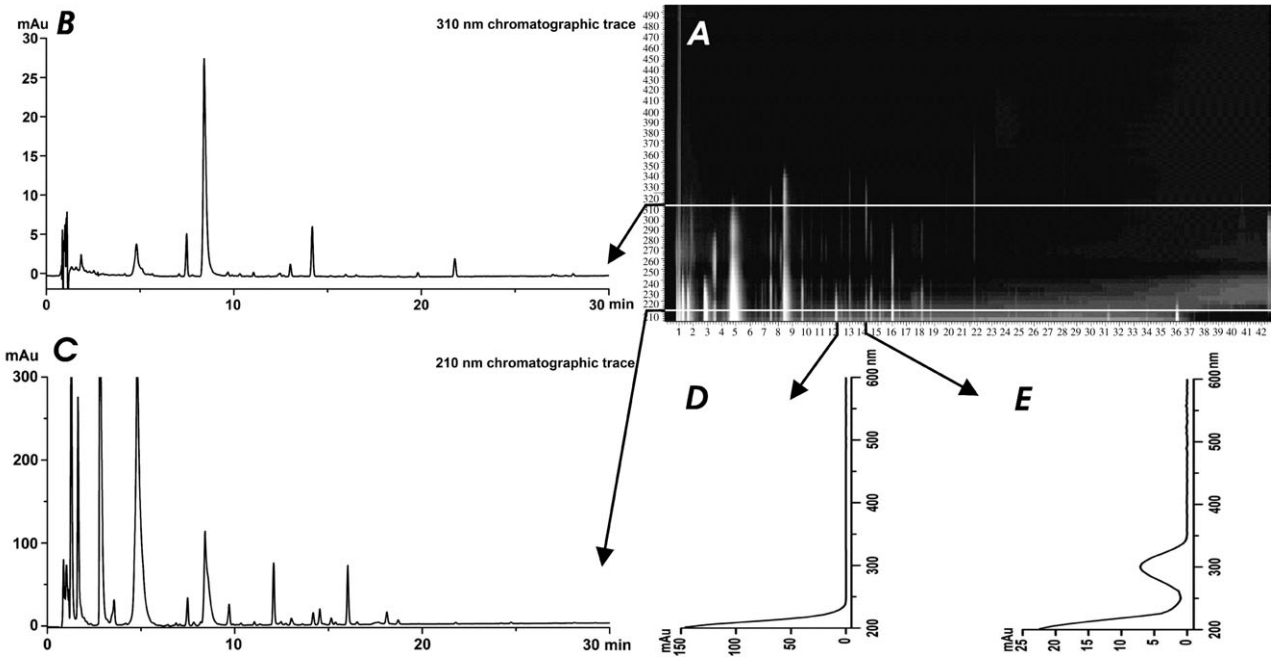

Fig. 3. The structure of HPLC-UV data matrix. Specific chromatographic traces (B, C) can be extracted from the data matrix in the retention time direction and spectral information for each time point can be found (D, E). The full data matrix can be viewed as an image (A) containing all information. From analysis of a *Penicillium polonicum* extract.

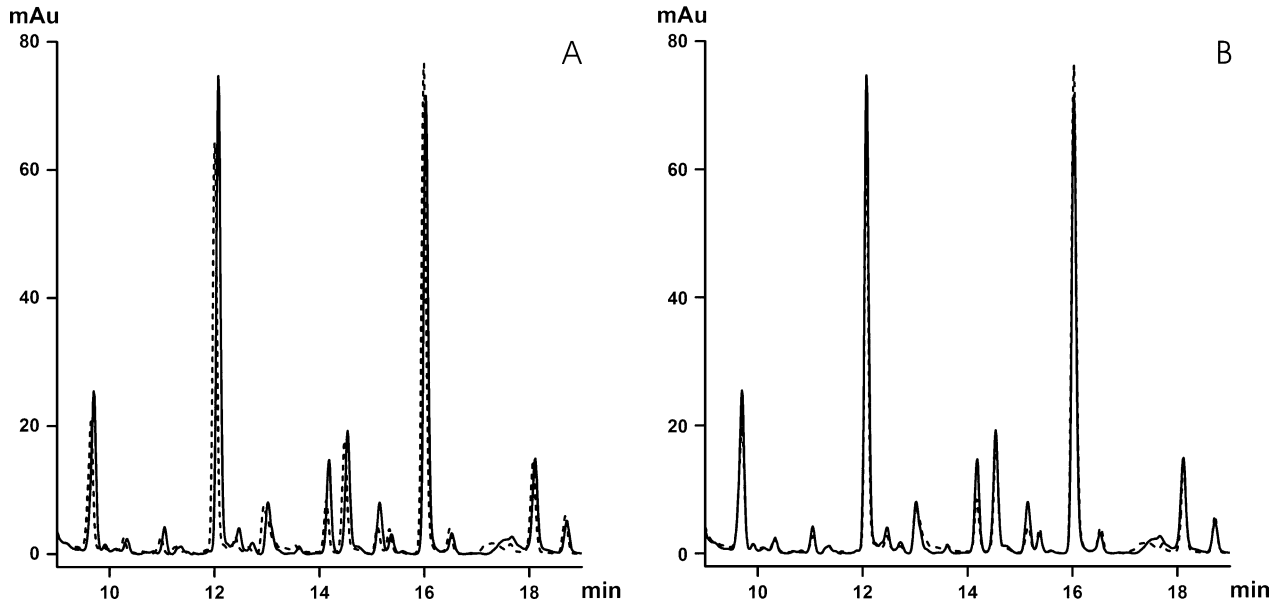

Fig. 4. Aligning of chromatograms from extracts of different *Penicillium polonicum* isolates by correlation optimized warping (cubed cophenetic correlation window 20 and slack 2; Nielsen *et al.* 1998). All information in the data matrices is used to align chromatograms, here the effects are shown as 210 nm chromatographic traces before aligning to the left and after aligning to the right.

are not classified together with the other isolates of the same species. In the extensive study by Nielsen *et al*. (1999) most species in the *Penicillium* subgenus *Penicillium* (212 isolates representing 41 species) were compared. They showed that more than 90% of the species could be classified correctly by CIA as judged by expert taxonomists using all the available taxonomic information. Furthermore, they illustrated that CIA has the potential to find sections of high/low similarity in the chromatograms (retention time domain), and thereby to identify peaks or compounds in common between samples or those that are unique and hence peaks or compounds that are responsible for separation of the species.

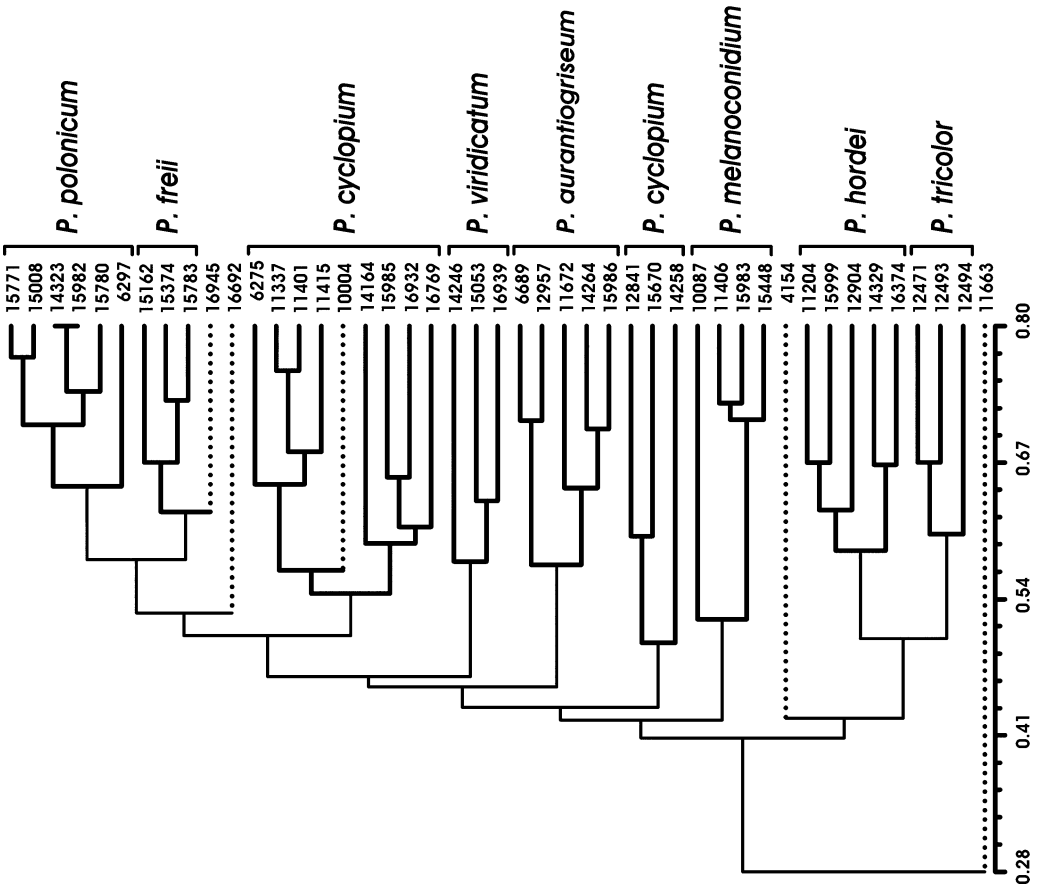

Fig. 5. Eight species from the series *Viridicata* in *Penicillium* subgenus *Penicillium* (5 or 6 isolates of each) were analysed by HPLC-UV. The resulting 45 chromatographic matrices (the images as shown in Fig. 3A) were aligned by COW and the similarities were calculated using the full data matrices (all about 3300 UV spectra in each file) to get an overall classification. In this case *Penicillium polonicum* IBT 15771 was used as the target for aligning (Nielsen *et al.*, 1998, 1999). The cluster analysis gives a clear grouping into species except *P. cyclopium* which, until recently, was split into two species (Frisvad and Samson, 2004). The isolate numbers refer to the IBT fungal collection held at BioCentrum-DTU, Technical University of Denmark.

Direct infusion electrospray mass spectrometry

While techniques based on HPLC analysis can give very detailed information it suffers from a rather long analysis time and the data are not easy to process automatically even by the alignment methods suggested above.

The commercial introduction of electrospray-mass spectrometry (ESI-MS) in the early 1990s opened the door to a whole new world of bioanalysis by mass spectrometry (Pramanik *et al.*, 2002). The most significant feature of ESI-MS is that it is a very soft ionization technique which, in many cases, will produce most the protonated molecular species (in positive ESI) for a broad range of different compounds with very high sensitivity. If this was the case, injection of complex samples would produce a mass profile of complex samples directly. In a proof-of-concept study, Smedsgaard and Frisvad (1996) showed that the most difficult to distinguish species from the series *Viridicata* (*Penicillium* subgenus *Penicillium*) could be classified within a few minutes by direct infusion mass spectrometry, diMS. Figure 6 shows an example of these early direct infusion mass profiles from three species that are very difficult to distinguish by traditional phenotypic taxonomy. Most predominant ions in these spectra correspond to the protonated mass of known and expected metabolites. While ion suppression is of concern in ESI-MS, the ion suppression does not seriously hamper the detection of expected metabolites although minor components may be lost.

This study was later extended to the entire *Penicillium* subgenus *Penicillium* where more than 75% of the species (47 at that time) could be classified correctly in accordance with expert classification by cluster analysis directly from the diMS spectra (Smedsgaard and Frisvad, 1997). The advantage of these nominal mass spectra is that they are easily aligned by unit mass binning and transfer into grid-like database structures. An obvious extension of a diMS approach is to store the spectra in a database using the database software included with most instruments. This will give a sample identification database rather than compound identification as anticipated by the manufactors. Smedsgaard (1997*b*) showed that about 75% of the species could be correctly retrieved by a cross-validation study using the simple software included with the mass spectrometer.

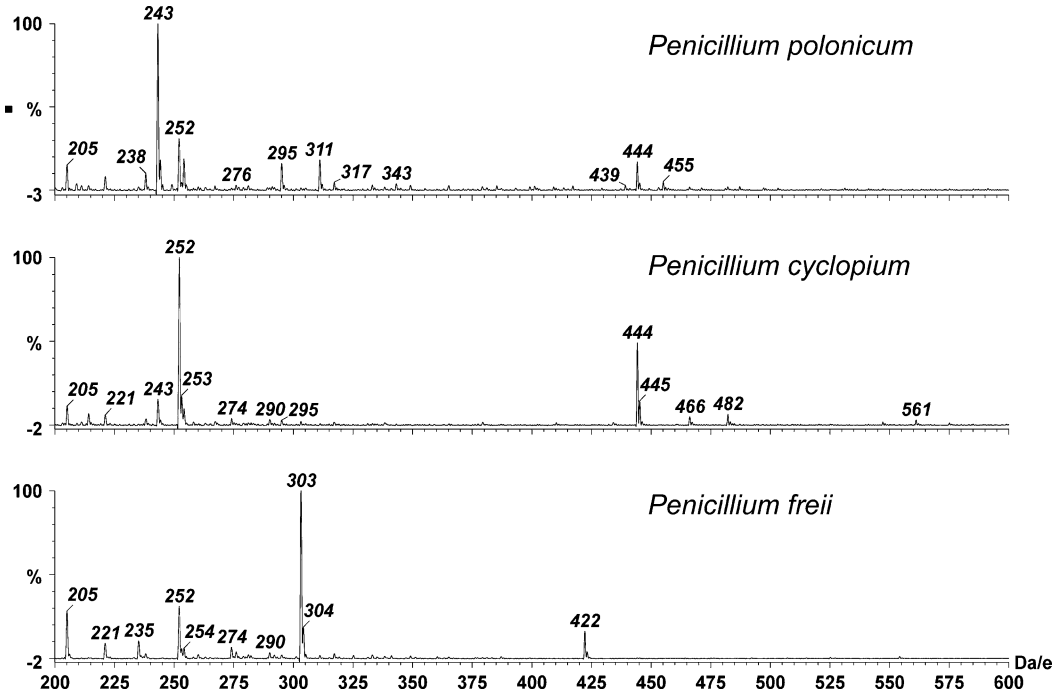

Fig. 6. Nominal mass spectra from direct infusion of fungal extract into electrospray mass spectrometry, diMS, shows the difference between three species which are difficult to distinguish by traditional phenotypic taxonomy.

With the arrival of relative easy-to-use and cost-efficient high resolution and high accuracy mass spectrometers, the usability of diMS for metabolite profiling has expanded dramatically. The major advantage is that each ion observed is much more likely to represent just one chemical formula and the accurate mass of this ion can be determined. This is illustrated in Fig. 7 where an accurate high resolution diMS spectrum of a *Penicillium polonicum* extract is shown.

To obtain maximum mass accuracy an internal mass reference is needed, in this case puberuline, a well-known metabolite produced by *Penicillium polonicum* (Frisvad *et al.*, 2004) is used as to correct the mass scale to give a mass accuracy that is typically better than 5 ppm. From the corrected mass spectrum it is possible to limit the number of possible candidates for each mass peak, thereby limiting the number of possible metabolites than can be attributed to each ion. From the accurate mass of the ions found in the spectrum in Fig. 7 a peak identity can be suggested using published metabolite data (Frisvad *et al.*, 2004). Table 1 compiles this information for *P. polonicum*, illustrating that ions corresponding to the protonated mass of many known metabolites can be found in these direct ESI-MS profiles.

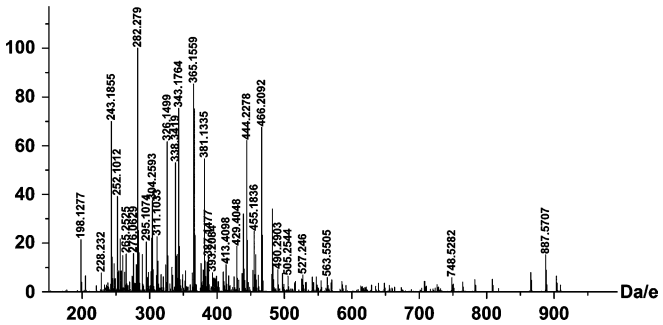

Fig. 7. Direct infusion mass spectrum of fungal extract using high resolution mass spectrometry, diMS, gives accuracy that allows the estimation of peak formula. Protonated puberuline ($M+H^+$ at 444.2287 Da/e) used as the internal mass references (see Table 1).

Accurate mass spectral processing

Nominal mass spectra are easily projected into a nominal mass grid (bins) as aligned integer mass variables without loss of information (Fig. 8). However, high resolution mass spectra require a more elaborate approach if the resolution and accuracy is to be maintained. This is illustrated in Fig. 9 showing a small section of nominal and high resolution mass spectra from the same sample. At nominal resolution, the four broad peaks are easily detected and binned. The high resolution data (about 8500 FWHM) show several more peaks and peak detection by centroiding, as normally used in mass spectrometry, produce a series of peaks. Binning these data without loss of information requires at least two decisions: the number of bins needed (thus the bin width) and what to do if more than one ion falls into each bin. Also, a very efficient smoothing and peak detection is needed to ensure that ions of the same mass are not artificially split into two centroid peaks due to poor peak detection and end up in different bins, blurring the

Table 1. *Calculation of the possible composition of ions found in the accurate mass spectrum showed in Fig. 7 from direct infusion mass spectrum of a* Penicillium polonicum *extract*

The proposed identification is based on metabolites known to be produced (Frisvad *et al.*, 2004). Elemental composition report: Multiple mass analysis: 93 mass(es) processed. Tolerance = 5.0 ppm DBE: min= −0.5, max=50.0. Elements: C<500 H<1000 N<7 O<13. Monoisotopic mass, odd and even electron ions. Intensity: 3.00% to 100.00%. 28700 formula(e) evaluated with 135 results within limits.

Meassured mass	Calculated mass	Error mDa	Error ppm	DBE	Formula $M+H^+$	Known metabolites of *P. polonicum*
Not found	153.0552				$C_8H_9O_3$	Orsellinic acid
Not found	171.0657				$C_8H_{11}O_4$	Penicillic acid
252.1013	252.1011	0.2	0.7	11.0	$C_{14}H_{12}N_4O$	3-meo-viridicatin
253.1052	253.1049	0.3	1.1	6.5	$C_9H_{13}N_6O_3$	
Not found	253.1063	−1.1	−4.2	6.0	$C_{11}H_{15}N_3O_4$	Viridicatol
279.1145	–					(dehydro-cyclopeptin)[a]
281.1293	281.1290	0.3	1.1	10.5	$C_{17}H_{17}N_2O_2$	Cyclopeptin
295.1085	295.1083	0.2	0.8	11.5	$C_{17}H_{15}N_2O_3$	Cyclopenin
298.1553	298.1556	−0.3	−0.8	9.5	$C_{17}H_{20}N_3O_2$	Leucyltryptophanyldiketopiperzine
311.1033	311.1032	0.1	0.4	11.5	$C_{17}H_{15}N_2O_4$	Cyclopenol
Not found	331.1770				$C_{17}H_{23}N_4O_3$	Anacine
334.1555	334.1556	−0.1	−0.2	12.5	$C_{20}H_{20}N_3O_2$	Rugulosuvine
401.2326	401.2328	−0.2	−0.5	8.5	$C_{24}H_{33}O_5$	Normethyl-verrucosidin
Not found	410.2444				$C_{24}H_{32}N_3O_3$	Verrucofortin
444.2287	444.2287	0.0	0.0	14.5	$C_{27}H_{30}N_3O_3$	Puberuline – LOCKMASS

[a] Not verified.

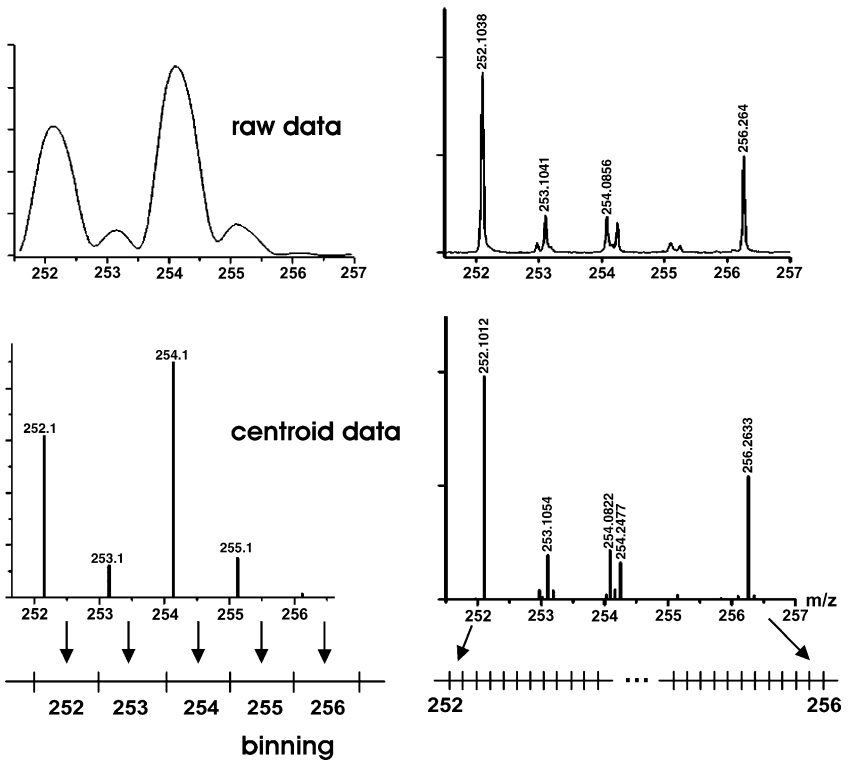

Fig. 8. Binning of mass spectral data. The nominal data to the left can easily be binned in nominal mass bins without loss of information, whereas the high resolution data to the right can not be binned as easily as the centroid masses are found on a continuous scale. This requires a decision to which bin a specific ion belongs.

downstream results. Finally, by using more bins, the number of empty bins will most likely increase and this will complicate the downstream processing and decrease the performance of, for example, principal component analysis and correspondence analysis. Binning mass spectra into one mass unit bin and subsequently using library searching is a straightforward process and performs well in many cases for high resolution data. In the case of high resolution mass spectra, much smaller bins can be used. However, rather than just using small bins one should always evaluate the data. In general, the decimal part of small biomolecules will be between the integer mass minus

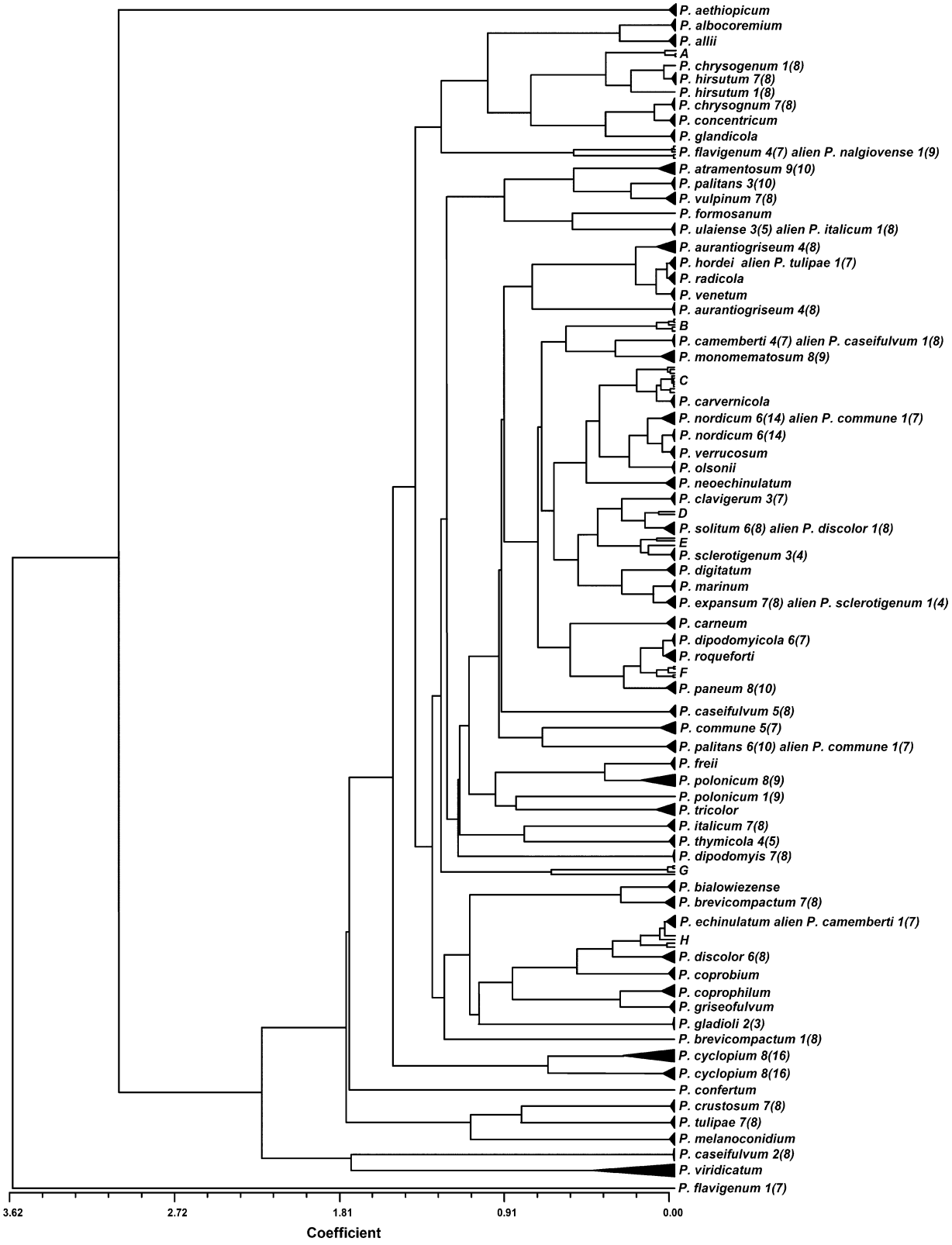

Fig. 9. Classification of *Penicillium* subgenus *Penicillium* species by AMS distance cluster analysis directly from accurate di-ESI mass spectra. Analysis includes: 491 isolates from 57 species. The number after the species is the number of isolates found in that cluster with the total number examined in brackets. The letters indicate mixed cluster (see Smedsgaard *et al.*, 2004; used with permission).

0.1 Da and plus 0.4 Da (below mass 1000 Da and consisting of only C, H, N, and O which accounts for most metabolites). Therefore, an intelligent binning system should take the average composition of biomolecules into account by dividing each nominal mass, m, into two bins: [m −0.1 Da <bin_1 ≤m+0.5 Da] and [m+0.5 Da < bin_2 ≤m+0.9 Da]. Of course several metabolites will fall into each bin as seen in the right part of Fig. 9, but this binning approach is simple and will decrease noise as dimers, multiple charged, and clusters will often fall into bin_2 which can be omitted in further data processing. Using this approach good performance has been obtained in the classification of fungi, whereas using narrow bins do not improve the results (J Smedsgaard, J Nielsen, unpublished results).

To perform automatic processing of high resolution mass spectra and maintain the full quality of the data without binning, Hansen and Smedsgaard (2004) have developed the concept of Accurate Mass Spectral (AMS) distance that avoids the need for binning. The core idea is to compare spectra pair-wise, establishing a series of unique ion-to-ion relations. From these unique relations it is possible to calculate the overall spectral distance taking both resolution and accuracy (if known) into account. Using this technique the species in the latest revision of the *Pencillium* subgenus *Penicillium* was classified automatically from high resolution mass spectra. 491 isolates was cultivated on two different media. All cultures were analysed by the plug extraction method (Smedsgaard, 1997*a*) followed by direct infusion electrospray high resolution mass spectrometry in positive mode. 200 to 600 ion peaks were detected in each spectrum after filtering. A metabolite known to be produced by each species was used for internal mass correction in the peak detection, centroiding process. The AMS distance was calculated (Hansen and Smedsgaard, 2004; Smedsgaard *et al.*, 2004) between all spectra obtained from each media (491) and showed that about 60–75% of the current 58 species were classified correctly according to the latest taxonomy by Frisvad and Samson (2004). The complete dendrogram from one of these analyses in Fig. 9 shows quite clear and well separated classes based on metabolite production as determined by diMS analysis.

Intelligent screening strategy

Combining the analytical techniques already discussed, the data processing method and biodiversity information has led to the formulation of an Intelligent Screening Strategy (ISS) as illustrated in Fig. 10 for the efficient discovery

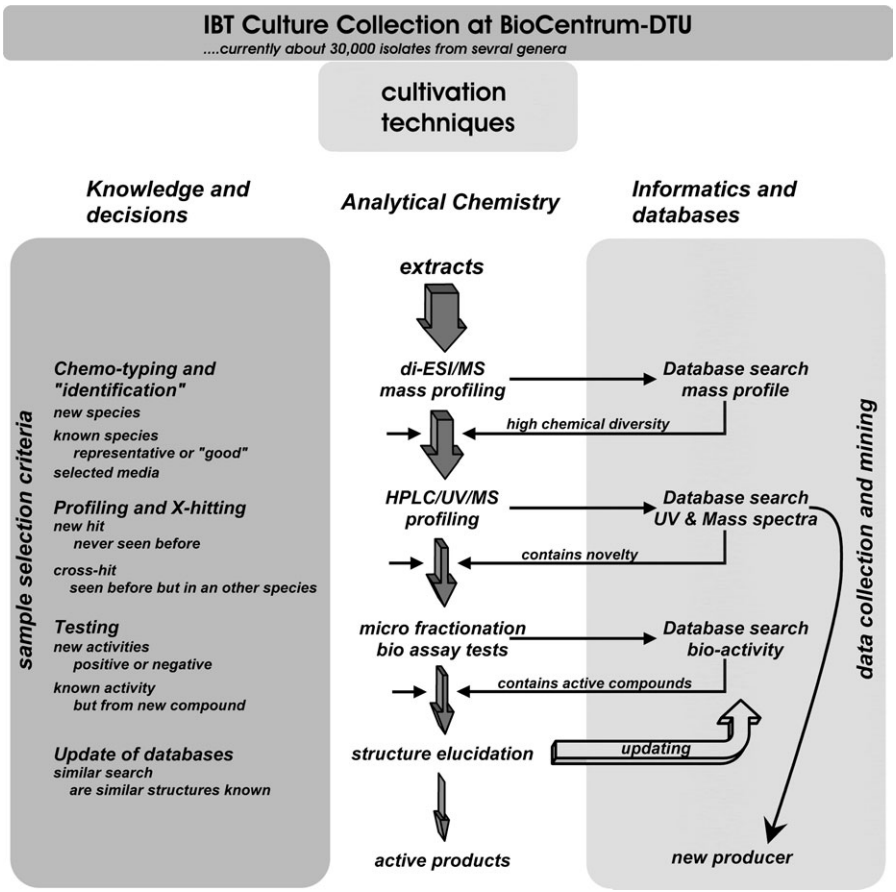

Fig. 10. The intelligent screening strategy, ISS, integrates biodiversity, systematics and chemical analysis with informatics and knowledge-based decisions for efficient screening of new usable compounds and organisms.

of microbial products and the exploitation of micro-organisms. The rationale behind ISS is that by efficient integration of biodiversity knowledge, systematics, chemical analysis, informatics, and databases allows a more efficient discovery process by limiting the more detailed studies to a relatively few representative isolates from each species (group) rather than by just studying anything that will grow. Knowledge about biodiversity and classification as illustrated by the principles of the *conceptual species model* is crucial to ISS.

From culture collections, isolates and species representing different habitats are selected and cultivated using a range of conditions (media and incubation conditions). Using high through-put extraction procedures (Smedsgaard, 1997*a*) crude extracts can be prepared rapidly. Automated diMS are used for an initial chemical classification and identification (using databases). From these results the chemically similar isolates are grouped and representative isolates are selected to proceed with the maximum chemical diversity. In the same process a pre-selection of cultivation media is done, again focusing on chemical diversity. The selected samples are then analysed by more elaborate techniques (typical LC-UV-MS) to de-screening for already known metabolites, to find compounds that have been seen before, but in other species and, finally, to find new compounds or those that are similar to known compounds. Only samples containing either novel compounds or that represent a new producer will be processed further. Importantly, all compound information is processed and appended to databases. At each step the number of samples processed is reduced, focusing on samples with maximum novelty. Samples containing chemical novelty will be subjected to micro-fractionation and can be tested in various bioassays. In the case of interesting activity, the structure of the compound(s) in these fractions will be elucidated. The key outcome is that ISS limits the number of samples subjected to the most complex analysis, while de-screening of samples avoids rediscovery of known metabolites and retaining of information. The ISS approach will very efficiently catch similar structures (analogues to known metabolites) or a new producing organism. In parallel, the ecological features will be examined by the integration of several cultivation conditions based on the ecological knowledge. Using ISS, Frisvad and co-workers have recently discovered and patented a series of compound with structures similar to the statins which include some of the most-used cholesterol-lowering agents (Frisvad *et al.*, 2003).

Functional genomics of yeast

The yeast *Saccharomyces cerevisiae* is amongst the most studied organisms and was the first eukaryote to have its genome sequenced. Large databases of experimental data are available for this organism, both at the transcriptional level and at the proteome level. Besides its wide application in biotechnology, this yeast represents a good model organism for studying molecular mechanisms in eukaryotic cells, and it is planned to use this organism for a concentrated effort in systems biology (www.yeastsystemsbiology.org). Recently, metabolite profiling was used for functional analysis of yeast mutants (Raamsdonk *et al.*, 2001; Allen *et al.*, 2003). Through analysis of the medium components taken up and secreted by the yeast cells (so-called footprinting) it was possible to cluster different yeast mutants and hereby obtain an insight into the function of specific genes. In this work, specific metabolites were not analysed, and raw analytical data were applied. This complicates the data analysis, and the approach will only work when many mutants are available to span the grid of information needed for the cluster analysis. In cases where there is less information it is interesting to apply more targeted techniques where specific metabolites are analysed, as this kind of information can be directly related to specific parts of the metabolic network. Using methyl chloroformate derivatization of metabolites, Villas-Bôas *et al*. (2003) developed a gas-chromatography mass-spectrometry method for the analysis of amino and non-amino organic acids in one single analysis. Their method enables analysis of up to 80 metabolites of the central carbon metabolism, and hereby it is possible to identify parts of the metabolism affected by specific mutations. This may open the way for the combination of metabolic models and metabolome analysis for a direct approach to functional genomics (Föster *et al.*, 2002).

Another approach for targeted analysis of the central carbon metabolism is the use of ^{13}C-labelled substrates and subsequent analysis of the incorporation patterns of the ^{13}C into the different intracellular metabolites. Analysis of the different isotopomers can be done by NMR or GC-MS (Nielsen, 2003), and based on the obtained data it is possible to gain an insight into the topology of the metabolic network, i.e. identify which metabolic routes are active at different growth conditions (Christensen and Nielsen, 1999). Through the combination of analysis of the incorporation patterns and metabolic models it is possible to quantify the fluxes through the different branches of the metabolic network (Gombert *et al.*, 2001). An example of a flux map for *S. cerevisiae* is shown in Fig. 11. The fluxes were estimated through feeding the cells with ^{13}C-labelled glucose, subsequent analysis of the isotopomers of the intracellular metabolites, and analysis of the experimental data through the use of a mathematical model of the metabolism. The fluxes are shown for two different mutants, and thereby it is possible to map the effect of specific mutations in the genome. In this case, the fluxes shown are for a wild-type yeast (in italics) and for a mutant with deletion of the gene *Grr1*, which is involved in glucose repression (Raghavendran *et al.*, 2004). In the wild-type strain there is glucose repression on respiration,

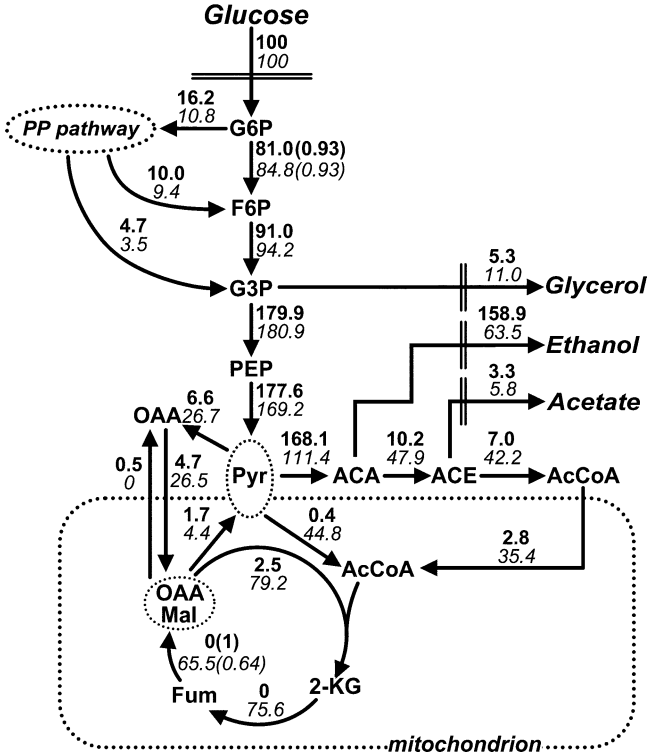

Fig. 11. Flux map for *Saccharomyces cerevisiae*. The fluxes were estimated through feeding the cells with ^{13}C-labelled glucose, analysis of the isotopomers of the intracellular metabolites, and analysis of the data using the mathematical model of the metabolism. In the wild-type strain there is glucose repression on respiration thus the flux through the TCA cycle is low. When *Grr1* is deleted there is a de-repression of respiration and the flux through the TCA cycle therefore increases.

and hence the flux through the TCA cycle is low. When *Grr1* is deleted there is a de-repression of respiration and the flux through the TCA cycle therefore increases. Using this approach it has been possible to gain insight into the function of the central carbon metabolism in different strains of the yeast *S. cerevisiae* (Gombert *et al.*, 2001; dos Santos *et al.*, 2003) and it has also been possible to perform a detailed functional analysis of specific enzymes/genes in the cell (dos Santos *et al.*, 2003).

Conclusions

Metabolite profiling is crucial for many aspects of both fundamental research and industrial application of filamentous fungi and yeast. In the microbial world the metabolites are of crucial importance in carrying much of the expressed functionality. In particular, micro-organisms from the Mycota are among the most important industrial micro-organisms and also have great impact on human life through food spoilage and as human pathogens. At the same time they constitute a huge biodiversity as they can be handled relatively easy in the laboratory. The general *Conceptual Species Model* presents a practical understanding of species features that, combined with the *Intelligent Screening Strategy* (ISS) give a solid platform to exploit and manipulate these micro-organisms through screening and metabolic engineering. It is interesting to study the metabolism of filamentous fungi and yeasts, in the context of functional genomics, metabolic engineering, and the discovery of metabolites. The overall strategy is equally applicable in many biotech areas besides fungi, for example, for the study and development of plants. These studies include both biodiversity, discovery of new compounds, bioprocesses, etc, but also more generalized studies of anything from different ecosystems to the single organisms. There is a large unexploited fungal biodiversity where intelligent screening opens the possibility (i) to recruit cell factories for the production of novel compounds that are otherwise difficult to produce by chemical synthesis, and (ii) the development and use of directed genetic modifications of cell factories, referred to as metabolic engineering (Nielsen, 2001), enables the development of novel, efficient and environmentally friendly bioprocesses. However, these studies will also improve the understanding of microbial ecology, for example, the ecology of mixed microbial cultures in foods and fungal infection in grain. It is, therefore, believed that the platform described here will play an important role in the trend towards development of green chemistry for the production of chemical compounds, with a substantial impact on society.

Supplementary material

Pictures of cultures grown on two different media and used for the dendogram in Fig. 5, are available at JXB online.

Acknowledgements

We thank the Danish Technical Research Council for support through the Program for Predictive Biotechnology (Functional biodiversity in *Penicillium* and *Aspergillus*) and the Centre for Advanced Engineering: The Centre for Microbial Biotechnology.

References

Agosti D. 2003. Encyclopedia of life: should species description equal gene sequence? *Trends in Ecology and Evolution* **18,** 273.

Allen J, Davey HM, Broadhurst D, Heald JK, Rowland SG, Oliver SG, Kell DB. 2003. High-through-put classification of yeast mutants for functional genomics using metabolic footprinting. *Nature Biotechnology* **21,** 692–696.

Blaxter M, Floyd R. 2003. Molecular taxonomics for biodiversity surveys: already a reality. *Trends in Ecology and Evolution* **18,** 268–269.

Christensen B, Nielsen J. 1999. Metabolic network analysis. *Advances in Biochemical Engineering/Biotechnology* **66,** 231.

Cole RJ, Cox RH. 1981. *Handbook of toxic fungal metabolites*. New York, USA: Academic Press.

Cole RJ, Jarvis BB, Schweikert MA. 2003. *Handbook of secondary fungal metabolites*, Vol. III. San Diego, USA: Academic Press.

Cole RJ, Schweikert MA. 2003*a*. *Handbook of secondary fungal metabolites*, Vol. I. San Diego, USA: Academic Press.

Cole RJ, Schweikert MA. 2003*b*. *Handbook of secondary fungal metabolites*, Vol. II. San Diego, USA: Academic Press.

Cuarro J, Gené J, Stchigel AM. 1999. Development in fungal taxonomy. *Clinical Microbiology Reviews* **12,** 454–500.

David H, Åkesson M, Nielsen J. 2003. Reconstruction of the central carbon metabolism of *Aspergillus niger*. *European Journal of Biochemistry* **270,** 4243–4253.

dos Santos MM, Gombert AK, Christensen B, Nielsen J. 2003. Identification of *in vivo* enzyme activities in the cometabolism of glucose and acetate by *Saccharomyces cerevisiae* by using ^{13}C-labeled substrates. *Eukaryotic Cell* **2,** 599–608.

Dunn CH. 2003. Keeping taxonomy based in morphology. *Trends in Ecology and Evolution* **18,** 270–271.

Famili I, Föster J, Nielsen J, Palsson BØ. 2003. *Saccharomyces cerevisiae* phenotypes can be predicted by using constrain-based analysis of a genome-scale reconstruction metabolic network. *Proceedings of the National Academy of Sciences, USA* **100,** 13134–13139.

Fiehn O. 2002. Metabolomics: the link between genotypes and phenotypes. *Plant Molecular Biology* **48,** 155–171.

Filtenborg O, Frisvad JC, Svendsen JA. 1983. Simple screening method for moulds producing intracellular mycotoxins in pure cultures. *Applied and Environmental Microbiology* **45,** 581–585.

Föster J, Famili I, Fu P, Palsson BØ, Nielsen J. 2003. Genmo-Scale reconstruction of the *Saccharomyces cerevisiae* metabolic network. *Genome Research* **13,** 244–253.

Föster J, Gombert AK, Nielsen J. 2002. A functional genomics approach using metabolics and *in silico* pathway analysis. *Biotechnology and Bioengineering* **79,** 703–712.

Frisvad JC. 1998. *Secondary metabolites and species models in* Penicillium *and* Aspergillus. Department of Biotechnology (now BioCentrum-DTU). Technical University of Denmark.

Frisvad JC, Filtenborg O. 1989. Terverticillate Penicillia: chemotaxonomy and mycotoxin production. *Mycologia* **81,** 837–861.

Frisvad JC, Lange L, Schnorr K, Stender S. Larsen TO. 2003. *Statin like compounds*. *World patent WO 03/048148 A2* filed 6–12–2003 in Denmark.

Frisvad JC, Samson RA. 2004. Polyphasic taxonomy of *Penicillium* subgenus *Penicillium*. A guide to identification of the food and airborne terverticillate Penicillia and their mycotoxins. *Studies in Mycology* **49,** 1–173.

Frisvad JC, Smedsgaard J, Larsen TO, Samson RA. 2004. Mycotoxins, drugs and other extrolytes produced by species in *Penicillium* subgenus *Penicillium*. *Studies in Mycology* **49,** 201–241.

Gombert AK, dos Santos MM, Christensen B, Nielsen J. 2001. Network identification and flux quantification in central metabolism of *Saccharomyces cerevisiae* under different conditions of glucose repression. *Journal of Bacteriology* **183,** 1441–1451.

Hansen ME, Smedsgaard J. 2004. A new matching algorithm for high resolution mass spectra. *Journal of the American Society of Mass Spectrometry* **15,** 1173–1180.

Harrigan GG, Goodacre R. (eds) 2003. *Metabolic profiling: its role in biomarker discovery and gene function analysis*. Boston, USA: Kluwer Academic Publisher.

Hofmann G, McIntyre M, Nielsen J. 2003. Fungal genomics beyond *Saccharomyces cerevisiae*? *Current Opinion in Biotechnology* **14,** 226–231.

Kell DB. 2004. Metabolomics and system biology: making sense of the soup. *Current Opinion in Biotechnology* **7,** 296–307.

Newman DJ, Cragg GM, Snader KM. 2003. Natural products as a source of new drugs over the period 1981–2002. *Journal of Natural Products* **66,** 1022–1037.

Nielsen J. 2001. Metabolic engineering. *Applied Microbiology and Biotechnology* **55,** 263–283.

Nielsen J. 2003. Guest commentary: it is all about metabolic fluxes. *Journal of Bacteriology* **185,** 7031–7035.

Nielsen KF, Smedsgaard J. 2003. Fungal metabolite screening: database of 474 mycotoxins and fungal metabolites for dereplication by standardized liquid chromatography-UV-mass spectrometry methodology. *Journal of Chromatography A* **1002,** 111–136.

Nielsen KF, Smedsgaard J, Larsen TO, Lund F, Thrane U. Frisvad JC. 2004. Chemical identification of fungi: metabolite profiling and metabolomics. In: Arora DK, ed. *Fungal biotechnology in agricultural, food, and environmental application*. New York: Marcel Dekker, 19–35.

Nielsen N-PV, Carstensen JM, Smedsgaard J. 1998. Aligning of single and multiple wavelength chromatographic profiles for chemometric data analysis using correlation optimized warping. *Journal of Chromatography A* **805,** 17–35.

Nielsen N-PV, Smedsgaard J, Frisvad JC. 1999. Full second-order chromatographic/spectrometric data matrices for automated sample identification and componenet analysis by non-data-reducing image analysis. *Analytical Chemistry* **71,** 727–735.

Pramanik BN, Ganguly AK, Gross ML. 2002. *Applied electrospray mass spectrometry*. New York, USA: Marcel Dekker.

Raamsdonk LM, Teusink B, Broadhurst D, *et al*. 2001. A functional genomics strategy that uses metabolome data to reveal the phenotype of silent mutations. *Nature Biotechnology* **19,** 45–50.

Raghavendran V, Gombert AK, Christensen B, Kötter P, Nielsen J. 2004. Phenotypic characterization of glucose repression mutants of *Saccharomyces cerevisiae* using experiments with ^{13}C-labeled glucose, *Yeast* **21,** 768–779.

Samson RA, Seifert KA, Kuijpers FA, Houbraken JAMP, Frisvad JC. 2004 Phylogentic analysis of *Penicillium* subgenus *Penicillium* using partial β-tubulin sequences. *Studies in Mycology* **49,** 175–200.

Smedsgaard J. 1997*a*. Micro-scale extraction procedure for standardized screening of fungal metabolite production in cultures. *Journal of Chromatography A* **760,** 264–270.

Smedsgaard J. 1997*b*. Terverticillate penicillia studied by direct electrospray mass spectrometric profiling of crude extracts. II. Database and identification. *Biochemical Systematics and Ecology* **25,** 65–71.

Smedsgaard J, Frisvad JC. 1996. Using direct electrospray mass spectrometry in taxonomy and secondary metabolite profiling of crude fungal extracts. *Journal of Microbiological Methods* **25,** 5–17.

Smedsgaard J, Frisvad JC. 1997. Terverticillate penicillia studied by direct electrospray mass spectrometric profiling of crude extracts. I. Chemosystematics. *Biochemical Systematics and Ecology* **25,** 51–64.

Smedsgaard J, Hansen ME, Frisvad JC. 2004. Classification of terverticillate *Penicillia* by electrospray mass spectrometric profiling. *Studies in Mycology* **49,** 243–251.

Smith JE, Solomons GL. 1994. *Mycotoxins in human nutrition and health report*. European Commission, Directorate-General XII, Science, Research and Development: EUR 16048 EN.

Stephanopoulos GN, Aristidou AA, Nielsen J. 1998. *Metabolic engineering*. San Diego, USA: Academic Press.

Sumner LW, Mendes P, Dixon RA. 2003. Plant metabolomics: large-scale phytochemistry in the functional genomics area. *Phytochemistry* **62,** 817–836.

Svendsen A, Frisvad JC. 1994. A chemotaxonomic study of the terverticillate penicillia based on high performance liquid chromatography of secondary metabolites. *Mycological Research* **98,** 1317–1328.

Tautz D, Arctander P, Minelli A, Thomas RH, Vogler AP. 2003. A plea for DNA taxonomy. *Trends in Ecology and Evolution* **18,** 70–74.

Taylor JW, Jacobsen DJ, Kroken S, Kasuga T, Geiser DM, Hibbert DS, Fisher MS. 2000. Phylogenetic species recognition and species concept in fungi. *Fungal Genetics and Biology* **31,** 21–32.

Turner WB. 1971 *Fungal metabolites*. London, UK: Academic Press.

Turner WB, Aldridge DC. 1983. *Fungal metabolites,* II. London, UK: Academic Press.

Villas-Bôas SG, Delicado DG, Åkersson M, Nielsen J. 2003. Simultaneous analysis of amino and non-amino organic acids as methyl chloroforomate derivatives using gas chromatography-mass spectrometry. *Analytical Biochemistry* **322,** 134–138.

Villas-Bôas SG, Mas S, Smedsgaard J, Nielsen J. 2004. Metabolome analysis by mass spectrometry. *Mass Spectrometry Reviews* (in press).