Available online at www.sciencedirect.com
SCIENCE DIRECT®
Journal of Chromatography B, 787 (2003) 129–148
JOURNAL OF CHROMATOGRAPHY B
www.elsevier.com/locate/chromb

Review

Analysis of the *Candida albicans* proteome II. Protein information technology on the Net (update 2002)

Aida Pitarch[a], Miguel Sánchez[b], César Nombela[a], Concha Gil[a,*]

[a]*Departamento de Microbiología II, Facultad de Farmacia, Universidad Complutense, Plaza Ramón y Cajal, s/n, 28040 Madrid, Spain*
[b]*Departamento de Microbiología y Genética IMB-CSCIC, Universidad de Salamanca, Salamanca, Spain*

Abstract

Candida albicans is an important fungal model organism of noteworthy clinical interest in modern medicine. Different initiatives addressing its sequencing and physical mapping have been carried out. The *C. albicans* genome sequence is currently near to completion at Stanford University, heralding new challenges in proteomic research and functional analyses of its gene products. This review presents an update of the most relevant data resources that are available through the World Wide Web to scientists working in the area of the analysis of the *C. albicans* proteome. An overview of the current status of the main universal protein sequence databases and specialized data collections for *C. albicans* is given. Various issues of the single public *C. albicans* 2D-PAGE database are also described, highlighting the significance of setting up graphical query interface-based databanks to visualize 2D-PAGE images through the Net. Finally, we also emphasize the pressing need to create a ''cyber-bioknowledge library'' that will integrate all the databases developed at the different levels for the understanding of life processes as well as bioinformatic tools for interpreting this deluge of data generated through the Internet.

Keywords: Reviews; Proteomics; *Candida albicans*; Protein databases; World Wide Web

Contents

*Corresponding author. Fax: +39-91-394-1745.
E-mail address: conchagil@farm.ucm.es (C. Gil).

1570-0232/02/$ – see front matter
doi:10.1016/S1570-0232(02)00762-6

1. Introduction

1.1. Candida albicans, a model organism of biomedical interest

The polymorphic fungus *Candida albicans* is an outstanding eukaryotic model organism for studying cellular differentiation and/or morphogenesis because of its ability to grow as different cellular forms and to switch among them under diverse cues [1,2]. Beyond its biological importance, and perhaps owing to its privileged condition of an opportunistic pathogen, the understanding of its putative pathogenicity and virulence factors still continues to be a major challenge to modern medicine. Its alarming incidence, particularly among the recent growing immunocompromised population (such as cancer patients, HIV-infected individuals and transplant recipients, to name but a few), where it can be life-threatening, combined with the limitations concerning the efficacy and side-effects of currently available drugs, the appearance of antifungal resistance, and the lack of specific and rapid diagnostic strategies, have undoubtedly given *C. albicans* research a noteworthy boost in the last two decades [3–6].

Given the biological and clinical relevance of this fungal organism, both public and private institutions have mainly focused their attention on the mechanisms involved in its morphogenesis, pathogenesis, virulence and drug resistance, as well as on the search for new molecular targets for antifungal drugs. In an attempt to reach accurate clues to all these topics, several approaches directed at the gene and protein levels (genomics and proteomics, respectively) have been developed (reviewed in [7–10]). There is no doubt, however, that the foreseeable annotation of the complete *C. albicans* genome sequence (see below), and eventually the development of functional genomics-based assays (gene–function–oriented analyses), will almost certainly provide significant breakthroughs and new insights into these inquiries.

Relevant information related to *C. albicans* genetics, molecular biology, molecular epidemiology and genome sequencing project, among others, is available on the Alces Web server at Uniform Resource Locator (URL) address http://alces.med.umn.edu/Candida.html.

1.2. Towards the forthcoming completion of the C. albicans genome sequence

Since the public release of the first entirely sequenced genome of a free-living organism in 1995 (that of *Haemophilus influenzae* [11]) to the present, the life science community has witnessed the completion of several genome sequencing projects – for many microorganisms and some higher organisms – at an exponential rate, demarcating a new era in the new millennium: i.e., the post-genome era. By far, one of the most exciting and ambitious projects has been the mapping of the whole human genome, its first draft being announced last year [12,13].

The genomes of two model yeast species have already been reported as fully sequenced, i.e., that of *Saccharomyces cerevisiae* (the first eukaryotic sequenced and annotated genome) [14] and that of *Schizosaccharomyces pombe* [15]. In contrast, *C. albicans* DNA sequencing is currently close to being

completed. Likewise, above 45 sequencing projects for important fungal model organisms (such as different *Aspergillus* species, *Cryptococcus neoformans*, *Coccidioides immitis*, *Fusarium sporotrichioides*, *Histoplasma capsulatum*, *Neurospora crassa*, *Pneumocystis carinii*, *Trichoderma reesei*, etc.) are still under way and will probably be concluded within a few years. The status of microbial genome sequencing projects can be found either at the Web site ''GOLD'' (Genomes OnLine Database; http://wit.integratedgenomics.com/GOLD) or ''TIGR (The Institute of Genomic Research) Microbial Database'' (http://www.tigr.org/), or alternatively on the NCBI (National Center for Biotechnology Information) Entrez server (http://www.ncbi.nlm.nih.gov/).

Regarding the human fungal pathogen *C. albicans*, diverse public and private initiatives aimed at its sequencing and physical mapping have been performed. Table 1 summarizes an updated list of these initiatives. Two *C. albicans* strains have been selected as paradigms (i.e., strain SC5314, a clinical isolate [16], and strain 1161, derived from 1006 [17]).

Accordingly, a *C. albicans* Genome Sequencing Project is being under way at the Stanford Genome Technology Center (Stanford University, US) with the support of the NIDR (National Institute of Dental and Craniofacial Research) and the Burroughs Wellcome Fund since October 1996 (http://www-sequence.stanford.edu/group/candida). Sequences were mainly obtained by shotgun sequencing of the whole genome of *C. albicans* strain SC5314. In May 2002, the assembly of the diploid genome shotgun sequence (assembly 19) was completed, this being the first public release of sequences for the two alleles of this human fungal pathogen. All sequence data will soon be announced in preliminary draft form – with possible errors and contamination from other species.

Table 1
WWW links relevant to the field of *C. albicans* research

Item	Name	Organization	URL	Contents
General	*C. albicans* information	Minnesota University	http://alces.med.umn.edu/Candida.html	Genetics, physical map, sequence data, strains and methods.
	WWW Virtual Library	Stanford University	http://genome-www.stanford.edu/	Links to the main Web sites related to data on *C. albicans*.
	The *Candida* Page		http://www.candidapage.com/	Links to Web sites related to *Candida* and candidiasis.
Genome sequencing projects	*C. albicans* genome sequencing project	The Stanford Genome Technology Center	http://www-sequence.stanford.edu/group/candida/	Sequencing of the whole genome of *C. albicans* strain SC5314.
	C. albicans physical map of chromosomes	Minnesota University	http://alces.med.umn.edu/Candida.html	Physical mapping of the chromosomes of *C. albicans* strain 1161.
	C. albicans genome sequencing project	The Sanger Institute and Aberdeen University	http://www.sanger.ac.uk/Projects/C_albicans/	Pilot sequencing project on *C. albicans* strain 1161.
	Private *C. albicans* sequence project	Genome Therapeutics Corporation	http://www.genomecorp.com/programs/pathogenome.shtml	Private initiative for *C. albicans* sequencing.
Genome databases	CandidaDB	Galar Fungail European Consortium	http://genolist.pasteur.fr/CandidaDB	Relational database for *C. albicans* genome.
	PathoGenome™ DataBase	Genome Therapeutics Corporation	http://www.genomecorp.com/programs/pathogenome.shtml	Commercial genomic database of *C. albicans*.
Proteome databases	MycoPathPD™ database	Incyte Genomics, Inc.	http://www.incyte.com/sequence/proteome/databases/MycoPathPD.shtml	Annotated bioknowledgebase for *C. albicans* proteins.
	COMPLUYEAST-2DPAGE database	Complutense University of Madrid	http://www.expasy.ch/ch2d/2d-index.html	*C. albicans* 2-DE maps with interactive identified proteins. Links to SWISS-Prot database.

In view of this, both the verification of its sequencing and the complete annotation of its sequence are expected to be published in a peer-reviewed journal in the foreseeable future. To address this concern, a *Candida* Annotation Working Group was formed at a recent genome workshop – held during the Sixth ASM (American Society for Microbiology) Conference on *Candida* and Candidiasis, in Tampa, Florida (13–17 January 2002) – with the aim of standardizing all *Candida* genomic information from all the different sources available. Currently, a high percentage of its sequence is already available in databases, which can be accessed through the Net by the *C. albicans* community (see Section 2.2).

The *C. albicans* Mapping Project is another public sequencing initiative focused on the construction of a detailed and complete physical map of its chromosomes (of *C. albicans* strain 1161). This mapping endeavor facilitates accurate location of all the genes and/or the assignment of sequence contigs to chromosomes (complete sequence-tagged site contig maps), and is thus a helpful tool for genome sequencing projects [18,19]. Physical mapping is being carried out at Minnesota University (US) and is funded by the National Institute of Allergy and Infectious Diseases (http://alces.med.umn.edu/candida/).

Likewise, the Sanger Institute, in collaboration with Aberdeen University (Scotland, UK), has also undertaken another pilot sequencing project on *C. albicans* strain 1161, financed by Beowulf Genomics (http://www.sanger.ac.uk/Projects/C_albicans). In an effort to create a physical map and hunt for new genes, a set of ten cosmids is currently being sequenced [18].

As a result of its demand by pharmaceutical industries, in parallel a private genomic company (Genome Therapeutics Corporation) has also achieved a *C. albicans* sequencing initiative (http://www.genomecorp.com/programs/pathogenome.shtml) (see Section 2.2.3).

The *C. albicans* sequence data derived from these projects should certainly yield crucial implications for our knowledge of its biology. In so doing, both the sequencing and mapping of *C. albicans* genome are first steps towards (i) the identification of genes potentially involved in virulence, (ii) the definition of their roles, and/or (iii) the search for new drug targets. Therefore, these data could eventually represent the blueprint for proteome research (proteomics) and functional studies (functional genomics), which can be supported by means of further analyses of 3-D structure predictions, metabolism (metabolome), signaling pathways (signalome), protein–protein interactions (interactome), and so forth (Fig. 1).

In short, as the complete annotation of the *C. albicans* genome sequence is reached, the main goal of *C. albicans* research community will be to gradually turn its attention to the end product of its genome (i.e., its proteome) and to the functions of its gene products. It is therefore time to think about proteins, since these are the functional biomolecules involved in most processes of living cells.

1.3. The awakening of proteomics in C. albicans research

Proteomics provides an overview of the complex world of proteins within a cell, a tissue or in a fluid, under the influence of different expression conditions (e.g., the cell cycle, endogenous and exogenous factors, etc.) at a given time point. This dynamic science, as implied by the name, studies proteomes – the set of proteins from a genome, cell, tissue or fluid [20]. The classical proteomic approach commonly involves (i) separation of proteins by two-dimensional polyacrylamide gel electrophoresis (2D–PAGE), (ii) identification and characterization of the resolved

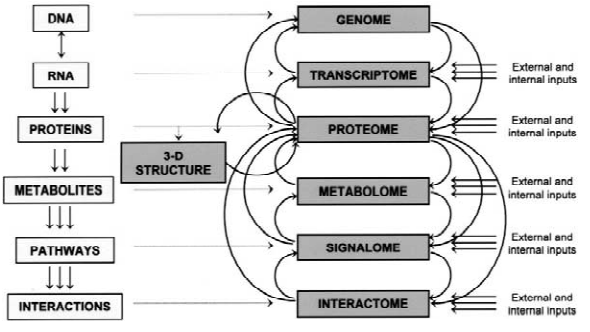

Fig. 1. Integration of the new biological ‘‘omes’’ terms. While genomes are static systems, the other ‘‘omes’’ are dynamic systems (they depend on external and internal inputs). Proteomics is the intermediate step to understanding the cellular organization and biological complexity of an organism. This platform seeks to correlate genomic and transcriptomic information with protein characterization and biological function.

proteins by mass spectrometry (laser desorption or electrospray MS), and (iii) data analysis using specialized bioinformatic tools and accumulation of the information in databases (bioinformatics) [21] (Fig. 2).

Currently, the systematic characterization of the *C. albicans* proteome is still in its infancy, making its first moves towards the understanding of dimorphism, virulence factors, cell wall, host response and drug resistance. *C. albicans* researchers are now witnessing the awakening of this discipline in their inquiries. Nevertheless, for further coverage on proteomic analyses as applied to *C. albicans*, the reader is referred to two recent reviews on this topic [9,10].

The establishment of complete proteomes requires annotated DNA and/or protein sequence databases, these tools being fundamental for such studies (see below and next section). Over the past few years, the *S. cerevisiae* sequence data have significantly assisted in *C. albicans* proteome analysis due to the high degree of similarity among nearly all sequences from both these closely related organisms [22].

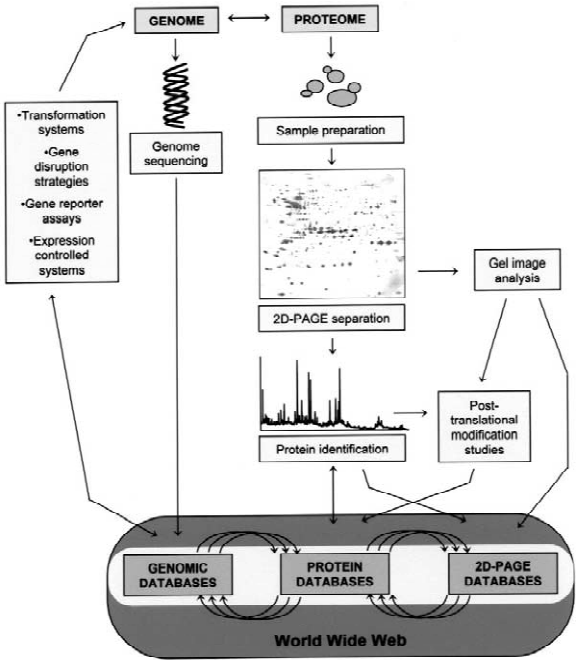

Fig. 2. Schematic representation of the classic proteomic approach and the basic bioinformatic building block for the study of proteomes. These databases are interconnected through the World Wide Web.

The field of bioinformatics thus plays a key role in proteome research and its importance in the biological sciences will presumably snowball progressively in the post-genome era. Annotated protein sequence databanks and 2D-PAGE databases have undoubtedly become the basic bioinformatic building block of proteomic investigations. As described in the following section, both databases are often interconnected with each other and with genome sequencing databases through the World Wide Web (WWW or Web), facilitating direct access to coding sequences and subsequent further studies (Fig. 2).

1.4. World Wide Web-based technology in the post-genome era

In view of the rapid emergence of large-scale genome and proteome projects, huge amounts of data are being amassed exponentially. It is evident that not all of the information generated can be published in conventional printed scientific journals, and hence other means of communication and data storage are required. As expected, the data generated could of course be compiled and organized into computerized databases with specialized software to update, query and retrieve data easily at any time. The aim of these databanks is therefore to store extensive information (such as nucleotide and amino acid sequences, 2-DE maps, biological literature, etc.) and ultimately, to facilitate their entire distribution.

An exciting milestone in the field of bioinformatics has been the recent development of Web-based technology. The World Wide Web has mainly allowed scientists around the world to (i) interface and integrate all these databases efficiently, and (ii) access readily their data files, without prior need to download full copies of the remote data collections in order to enter into their content on local computers [23]. In a nutshell, WWW is currently very helpful for integrating, locating and exchanging knowledge throughout the world, with no audience restriction.

Web pages are special hypertext documents that have been designed in a simple programming language, HTML (HyperText Mark-up Language). HTML documents, displayed by a Web browser, are usually linked to others through active hypertext

cross-references. This enables users to gain more information about the issue referenced merely by clicking on it. In the same way, one database (HTML data) can consequently be cross-referenced to others, making it possible to query remotely located databases, navigate across them and/or retrieve pieces of information stored within the system using different criteria. Indeed, these active hypertext links constitute the core component of database integration [24] (see Section 3.2.). In addition, user-friendly interfaces are often developed to facilitate the search for or recovery of data of interest.

The WWW paradigm has undoubtedly revolutionized data dissemination among the scientific community in recent years [25]. Thus, many Web sites related to life science research have already been constructed, and have become an essential part of the professional activities of the scientific community. A selection of several Internet Web sites that provide useful information associated directly with different issues of *C. albicans* and candidiasis (such as the *C. albicans* information server of Minnesota University, data on symptoms, diagnosis and treatment of Chronic Candidiasis Syndrome, etc.) can be reached, for instance, through links on ''The *Candida* Page'' and ''WWW Virtual Library'', which, in turn, can be accessed through the Net at URL address http://www.candidapage.com and http://genome-www.stanford.edu/, respectively. A list of handy Web sites containing the most relevant genomic and proteomic databases concerning this human fungal pathogen is shown in Table 1.

The goal of the present paper is to offer an overview of the current status of the most significant data resources available on the World Wide Web to researchers interested in the field of the analysis of the *C. albicans* proteome. Given the large number and variations in content of the existing data collections, here we focus mainly on two important types of databases for proteomics, i.e., annotated protein sequence and two-dimensional gel electrophoresis databases. Along the next chapter, we survey both the general protein databanks and specialized data collections for *C. albicans* currently available on the Internet. After this, in the subsequent section, we debate the implications of proteome databases in the post-genome era, and describe various aspects of the single *C. albicans* 2D-PAGE databank up to date (i.e., COMPLUYEAST-2DPAGE database). Finally, we offer our own personal view of future insights into the field of databases.

2. ''Surfing the protein networks''. Currently available sequence databases on *C. albicans*

Considering the diverse nature of the information generated in proteomic studies, it is clear that a broad variety of data resources is required for querying and/or retrieving all this information. For instance, databases for (i) protein sequences (e.g., SWISS-PROT, TrEMBL, PIR, GenPept), (ii) nucleotide sequences (e.g., EMBL, GenBank, DDBJ, dbEST), (iii) genetic organization (e.g., OMIM), (iv) sequence patterns and protein families (e.g., PROSITE, BLOCKS, Pfam, PRINTS), (v) 2D-PAGE (e.g., SWISS-2DPAGE), (vi) three-dimensional structures (e.g., PDB, HSSP), (vii) post-translational modifications (e.g., GlycoSuiteDB, CarbBank) and/or (viii) metabolic pathways (e.g., ENZYME, KEGG, WIT) can certainly be powerful bioinformatic tools for supporting proteome research [26,27] (Fig. 3).

In turn, the major public DNA and protein sequence databanks, along with other specialized data collections, are interconnected through the World Wide Web by means of an integrated database retrieval system (such as Entrez [28] and the SRS (Sequence Retrieval System) [29] network browser provided by the National Center for Biotechnology Information (NCBI) in Bethesda, US (http://www.ncbi.nlm.nih.gov) and the European Bioinformatics Institute (EBI) in Hinxton, UK (http://www.ebi.ac.uk), respectively). This enables a substantial enlargement of the knowledge library related to protein data. In this sense, for example, currently the sophisticated SWISS-PROT database is explicitly and implicitly cross-referenced with 45 and 23 different data collections, respectively (Release 40.24), playing a key role as the major nucleus of biomolecular database interconnectivity [30,31] (see below). In brief, therefore, the active integration of databanks over the Internet could offer an essential

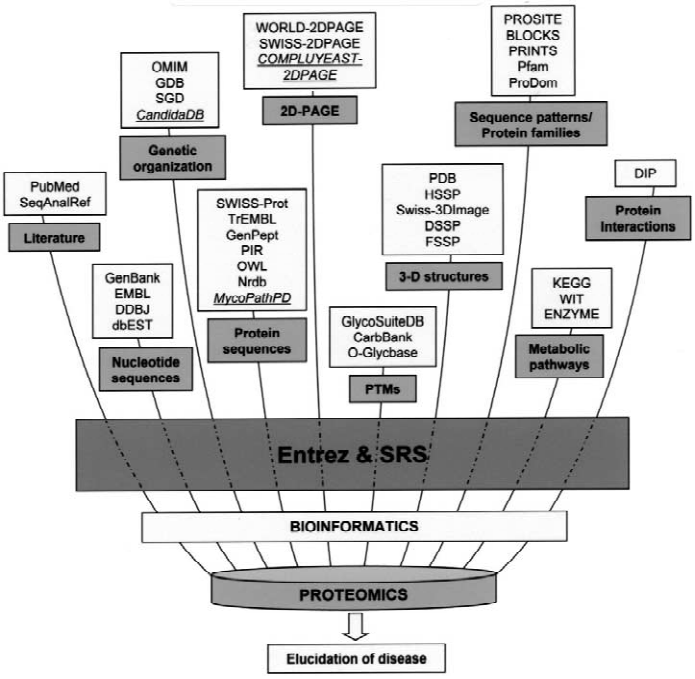

Fig. 3. Main WWW-available databases for the study of proteomes. Entrez and SRS are information search and retrieval systems for molecular biology databanks. PMFs, post-translational modifications.

breakthrough in our understanding of protein functions and of the complexity of biological systems.

Protein identification entails accessing sequence-related databases. However, there are three main classes of sequence databases available (i.e., nucleic acid sequences, protein sequences and protein tertiary structures) together with a wide variety of specialized data collections on the Web. Within the last years, protein databanks, unlike DNA sequence databases, have certainly undergone a linear growth [27]. Most sequence data collections essentially contain two categories of data for each sequence entry, i.e., (i) the core data (or sequence data for protein identification), and (ii) the annotation (for further valuable information about the protein in question). The annotation process is accomplished by external professionals – namely, annotators or curators – that extract and gather this information from (i) international journal articles reporting sequencing and/or characterization, (ii) review articles, (iii) patent applications or (iv) data directly submitted to the database by personal communications. Some of these WWW-accessible databanks are free but others require a subscription or the payment of a license fee for reaching their information.

In this section we merely concentrate on the most relevant sequence-related databases for *C. albicans*. Two levels of information resources are contemplated: (i) general databases (from many biological organisms, naturally including *C. albicans*) and (ii) specialized databases (from a particular species, i.e., *C. albicans*).

2.1. General protein sequence databases

Most universal protein sequence databases often derive from general DNA sequence databanks that have been directly translated (such as, SWISS-PROT/TrEMBL and GenPept data collections, which basically contain translations of the EMBL and GenBank nucleotide sequence databanks, distributed by EBI and NCBI, respectively).

Here we comprehensively describe both the SWISS-PROT protein knowledgebase and the

TrEMBL computer-annotated supplement to SWISS-PROT and we also briefly outline other protein sequence databases, highlighting their current status as regards *C. albicans* protein entries.

2.1.1. SWISS-PROT™ knowledgebase and TrEMBL™ supplement database. SPTR database: an updated, comprehensive, non-redundant protein sequence database

SWISS-PROT is an annotated universal protein sequence database maintained by the Swiss Institute of Bioinformatics (SIB) in Geneva (Switzerland) in collaboration with the EMBL Outstation – the European Bioinformatics Institute (EMBL/EBI) in Heidelberg (Germany) and in Hinxton (UK), respectively. The ExPASy (from Expert Protein Analysis System) WWW molecular biology server, or alternatively the EBI SRS server, allow users to access this databank through the Internet at URL address http://www.expasy.ch/sprot or http://www.ebi.ac.uk/sprot, respectively [30].

The fact that SWISS-PROT is regarded as a protein knowledgebase can basically be attributed to (i) its high-quality annotation (e.g., description of the function/s of the protein, its subcellular location, its post-translational modifications, its domains, sites and structures, its similarities to other proteins, its variants and other noteworthy issues), (ii) its minimum or non-redundancy (few duplicate copies) and (iii) its sophisticated integration with many other biomolecular databases (currently it is cross-referenced with about 60 different data resources through the user-friendly hypertext model) [32].

However, this detailed and labor-intensive annotation procedure is evidently slower than the arrival of new raw data from genome projects. It is thus essential to develop other types of data collection that amass the sequence information prior to being annotated so as to maintain a high-quality level [33]. The TrEMBL (Translation from EMBL) database was created precisely for this purpose; that is, as a supplement to SWISS-PROT. It comprises the translations of all the coding sequences present in the EMBL (European Molecular-Biology Laboratory) nucleotide sequence databanks [34] that (i) have not yet been incorporated into SWISS-PROT (so-called SpTrEMBL – from SWISS-PROT TrEMBL), and (ii) will not be included in SWISS-PROT, such as patent data, synthetic protein sequence, immunoglobulins, T-cell receptors, small fragments or other sequences not coding for real proteins (namely RemTrEMBL – from Remaining TrEMBL) [30]. In turn, the weekly update to TrEMBL from the new sequences deposited at the EMBL (EMBLnew) is compiled in TrEMBLnew.

A state-of-the-art view of the non-redundant protein sequences is currently given by the comprehensive SPTR database, which is accessible as the Swall databank on the EBI SRS server. This data collection consists of (i) SWISS-PROT, (ii) SpTrEMBL, and (iii) TrEMBLnew, promoting the creation of an updated, complete and non-redundant protein sequence database.

At present, approximately 113 000 annotated sequence entries from about 7500 different species have been reported in the SWISS-PROT knowledgebase. Currently, SWISS-PROT (Release 40.26) and TrEMBL (Release 21.7) contain 263 and 451 sequence entries from *C. albicans*, respectively. The SPTR (or SWall) database displays a total of 734 protein sequences from *C. albicans* (making an allowance for the updated 20 entries present in TrEMBLnew). Nevertheless, a few entries from SWISS-PROT include partial *C. albicans* protein sequences, corresponding to fragments of its gene products that have been submitted to the database, for instance, as sequence data from *N*-terminal micro-sequencing or peptide fragmentation by mass spectrometry (e.g., methionine synthase or METE_CANAL [22,35]).

Hopefully, *C. albicans* is one of the 18 model organisms with sequence entries available in SWISS-PROT that have been selected (i) to provide cross-references to specialized genetic databases (see below), specific indices and high level of annotations, as well as (ii) to promote its completion as soon as possible [34]. This will undoubtedly smooth the progress of forthcoming *C. albicans* proteomic researches.

2.1.2. Other public protein sequence information resources

Although the SWISS-PROT knowledgebase is one of the most important currently available protein databases, other protein sequence data collections are

also worthy of attention. Below we briefly describe some of them.

(i) *GenPept translated protein-coding sequence database*. GenPept is an automatic translation of the corresponding coding regions of sequences released by the GenBank database at NCBI (http://www.ncbi.nlm.nih.gov). About 1533 *C. albicans* protein entries are currently available.

(ii) *PIR-International Protein Sequence Database (PIR-PSD)*. The PIR-PSD is distributed by the PIR (Protein Information Resource) in collaboration with MIPS (Munich Information Center for Protein Sequences) and JIPID (Japan International Protein Information Database). This databank is a non-redundant, comprehensive, expert-annotated, fully classified and extensively cross-referenced protein sequence database [36] (http://pir.georgetown.edu/). It contains (i) sequence data derived from the translations of the GenBank, EMBL and DDBJ (DNA DataBank of Japan) nucleotide sequence databases, (ii) the superfamily/family-based classification of proteins (this being the main characteristic of the PIR-PSD), and (iii) further direct submission or literature-related data on each protein. The current release has 173 *C. albicans* entries (Release 73.03).

(iii) *OWL composite protein sequence database*. This data collection is a non-redundant combination of data from diverse primary sources, including (i) SWISS-PROT (which is the highest priority resource), (ii) GenBank (translation), (iii) PIR and (iv) PIR-NRL3D (from sequence and annotation in the Protein Data Bank (PDB) of three-dimensional structures), which can be accessed at http://www.bioinf.man.ac.uk/dbbrowser/OWL. To circumvent redundancy, both identical sequences and those differing only trivially are eliminated when protein entries from the latter three databases are compared against those from SWISS-PROT [37]. Currently, it contains 330 entries for *C. albicans*.

(iv) *Non-redundant protein database (Nrdb)*. Nrdb has been provided by NCBI as the default option for searching for sequence similarities using the Blast (Basic Local Alignment Search Tool) heuristic algorithm [38] (http://www.ncbi.nlm.nih.gov). This data resource is a composite of data from (i) SWISS-PROT (including updates), (ii) GenPept (and GenPept updates), (iii) PIR and (iv) PDB, the criterion for bypassing redundancy being based on the elimination of two closely matching sequences. However, it may generate certain shortcomings, such as the presence of several entries of the same protein that differ as a result of polymorphisms, sequencing errors, etc. Hence, Nrdb is a database that is updated most frequently, and is larger and less efficient to search than OWL.

(v) *Patent_prt library*. This databank includes patented protein sequences available in the public domain from the European Patent Office (http://www.ebi.ac.uk). The current version consists of around 360 proteins from *C. albicans*.

2.2. Specialized data collections

Here we survey the two major specialized annotated databases for *C. albicans*. Both of them offer a comprehensive dataset of DNA and/or protein sequences as well as remarkable literature-based annotations. Data sources mostly arise from the *C. albicans* Genome Sequencing Project accomplished at the Stanford Genome Technology Center (see Section 1.2.). Other data collections are also mentioned.

2.2.1. CandidaDB™, a relational database for the analysis of the C. albicans genome

CandidaDB is a *C. albicans* genomic database developed in collaboration by members of the Pasteur Institute (France) and the Galar Fungail – i.e., Fungal Disease in Gaelic – European Consortium (comprising the participation of eleven European laboratories) within the framework of the European project ''Novel approaches for the control of fungal disease'' (http://www.pasteur.fr/Galar_Fungail/CandidaDB/). It has also been supported by funds from the French Ministry of Research.

Currently, this genomic data collection is freely accessible to the *C. albicans* research community on the Web at URL address http://genolist.pasteur.fr/CandidaDB. In addition, a flat-file distribution of CandidaDB, which includes three series of files, is also reachable, it thus being possible to download the data contained in this server, either in EMBL format (i.e., the complete database) or in FASTA format (i.e., the protein-encoding DNA sequences or the protein sequences).

Data on contig DNA sequences of Assembly 6

(standard assembly) of the *C. albicans* genome available at the Stanford Genome Technology Center have been exploited to implement the CandidaDB. The strategy used in its creation involves (i) the identification of ORFs that encode proteins exceeding 150 amino acids, with homologues in other public data collections or whose coding probability is in accordance with GeneMark prediction, (ii) the detection of repeated protein coding sequences (owing to the heterozygous nature of *C. albicans*), and finally, (iii) the annotation of all ORFs or protein coding sequences.

This server contains (i) annotated and integrated DNA sequence data derived from Standford's *C. albicans* sequencing project (Assembly 6 of the *C. albicans* genome), (ii) further information from *C. albicans* entries displayed in EMBL/GenBank/DDBJ nucleotide sequence databases, (iii) relevant information on the genetic organization (e.g., gene location – position and chromosome – and physical and genetic distances between genes, etc.), (iv) fruitful data on the coding sequences and proteins (e.g., description of their functions, isoelectric points and molecular masses, homology with other species, etc.), and (v) cross-references to other nucleotide and protein sequence databases (such as the SGD (*Saccharomyces* Genome Database) and the YPD (Yeast Proteome Database)). In the light of this, this genomic databank certainly incorporates useful information for proteome studies.

The CandidaDB format is structured in a single Web page-per-protein. This offers information about the queried gene and protein (from the paradigm *C. albicans* strain SC5314), such as gene name, accession number, description and functional category, location and many other issues. Consequently, both navigation across the different database entries and recovery of information can be assisted easily using diverse criteria (i.e., gene names, location, keywords, etc.). In addition, owing to the user-friendly query interface exploited in this database, it is also easy to browse through it, and search and retrieve any type of data concerning the *C. albicans* genome.

Currently, the CandidaDB contains 6165 non-redundant sequence entries, which correspond to nearly 5920 genes of *C. albicans* strain SC5314. In agreement with haploid genome size relative to *C. albicans* (16 Mb), approximately 95% of its genes are already annotated in this relational database, although some mistakes and bugs may still remain. This vast data set could provide new insight into improvements in current antifungal therapies.

2.2.2. MycoPathPD™ (CalPD™), a pathogen fungal species-specific volume of the Proteome BioKnowledge Library

The Proteome BioKnowledge Library (Incyte Genomics, Inc., California, US) is a powerful integrated relational resource for proteomic researches (http://www.incyte.com). The available protein information of the model organism *C. albicans* was integrated into a specific volume of the BioKnowledge Library, called CalPD (*Candida albicans* Proteome Database), until about 3 years ago [39]. In 2001, however, CalPD was expanded into a unified pathogen fungal species-specific volume of this library, known as MycoPathPD (MycoPath Proteome Database), which also includes further *Candida* species (i.e., *C. dubliniensis*, *C. glabrata*, *C. guillermondii*, *C. krusei*, *C. lusitaniae*, *C. parapsilosis*, *C. pseudotropicalis*, *C. tropicalis*) as well as diverse important human fungal pathogens (i.e., *Aspergillus* species – *A. flavus*, *A. fumigatus*, *A. niger*, *Blastomyces dermatitidis*, *Coccidiodes immitis*, *Cryptococcus neoformans*, *Histoplasma capsulatum* and *Pneumocystis carinii*) [40]. Regardless of this, we only focus on *C. albicans* protein information in the rest of this section.

This volume can currently be found on the Web at http://www.incyte.com/sequence/proteome/databases/MycoPathPD.shtml, a license fee being required since June 2002, so that it is accessible to both academic and corporate users. Until not long ago, this knowledgebase was available to academic users upon registration.

MycoPathPD is a curated proteome database that compiles, organizes and displays all comprehensive knowledge (both full protein sequences and scientific literature-based data) available about each protein predicted from Stanford's *C. albicans* sequencing project and GenBank in a user-friendly Protein Report format (one page-per-protein) [40,41]. Each Protein Report, displaying information extracted and gathered from the scientific literature by qualified annotators (Ph.Ds with knowledge of the model organisms), can be divided into four sections: (i) an initial title line (one-to-two lines containing a brief and precise description of the updated protein func-

tion), (ii) an upper section (a table including the properties of that protein, such as its cellular role, biochemical function, subcellular localization, sequence data, related proteins, protein interactions, protein modifications, among others), (iii) a lower section (free-text annotations structured by topic headings, e.g., function, catalytic activity, pathway, genetic and physical interactions, regulation, gel mobility, functional genomics, etc.), and (iv) a final part, which is a comprehensive reference list cross-referenced with MEDLINE abstracts, i.e., the biomedical literature (PubMed, reachable using Entrez retrieval system). It has the same format as other volumes of the Proteome BioKnowledge Library, such as YPD (from *Saccharomyces cerevisiae*), WormPD (from *Caenorhabditis elegans*) or PombePD (from *Schizosaccharomyces pombe*) [39,40,42].

However, this bioknowledgebase could in theory contribute to the search for potential clues to the structure, function, subcellular location or interactions of any unknown protein from other species boundaries.

About 9850 entries for end gene products from *C. albicans* are presently available at MycoPathPD. Some of them correspond to predictions of ORFs generated in the genome-sequencing project of this human fungal pathogen.

2.2.3. A private C. albicans genomics databank

Since non-public institutions have accomplished private initiatives for the sequencing of the *C. albicans* genome, annotated sequence databanks have evidently been developed.

The PathoGenome™ DataBase (Genome Therapeutics Corporation (GTC), US) is a commercial resource of microbial genomic information (http://www.genomecorp.com). It includes a private and annotated *C. albicans* genomics database. To date, however, only important pharmaceutical companies have become subscribers to this data collection.

3. "Web surfing" across the *C. albicans* proteome. Two-dimensional gel electrophoresis databases of *C. albicans*

At present, high-resolution two-dimensional polyacrylamide gel electrophoresis (2D-PAGE) undoubtedly continues to be one of the most widely used tools for the simultaneous separation of complex protein mixtures or, more accurately, for the comprehensive analysis of entire proteomes [43]. The evaluation of protein spots on 2-DE gels can be assisted by means of diverse software packages for 2-D image analysis, such as Melanie, PDQuest, or Imagemaster [44,45]. In turn, each 2-D protein spot can also be characterized and matched to its corresponding gene sequence using mass spectrometry (MS) and different WWW-available software algorithms that correlate MS data with sequence databases, this currently being the strategy of choice for proteome identification [21,46].

Considering the extensive information afforded by these proteomic studies (e.g., reference 2-DE image maps, data on identified proteins – experimental isoelectric point (p*I*) and the molecular weight (Mw) of each protein spot, functions, quantitative analysis or relative abundance of these proteins (differential expression patterns), subcellular location, induction by specific stimuli, post-translational modifications, MS and MS/MS sequences, etc. – further literature-gathered annotations, to name but a few, it is clear that a rapid accumulation of 2-DE-related data will be generated in a short time. It is thus necessary to have available a specialized type of image-based databanks that (i) are, of course, accessible over the Internet to the 2-DE community, (ii) integrate and amass all the proteomic information obtained and, ultimately, (iii) use graphical query interfaces for the visualization and exploration of 2-DE gel images.

Fortunately, in the last few years two key factors have facilitated the setting up of these 2D-PAGE database servers on the Net, i.e. (i) the high intra- and inter-laboratory reproducibility of the available 2-DE technology – due mainly to the application of immobilized pH gradients (IPGs) for the first-dimension separation, and (ii) the recent introduction of more rapid and sensitive protein identification methods (e.g., MS) [47].

Several pioneering 2D-PAGE databases of biological and medical interest, although as yet uncompleted, have emerged on the World Wide Web network in recent years (e.g., SWISS-2DPAGE [48–50], as the first WWW-accessible 2-DE databank, and other locally maintained databases, containing reference 2-DE maps for biological samples from mammalia, plants, yeasts, bacteria, viruses, parasites

or cell lines). These 2-DE image data collections can be freely reached from the Web site WORLD-2DPAGE through the ExPASy WWW molecular biology server, which in turn can be accessed at URL address http://www.expasy.ch/ch2d/2d-index.html [51,52]. In addition, this server offers the possibility of exploiting a specialized software tool (2D Hunt) when searching for 2-DE related sites on the Web (i.e., both 2-DE databases and documents), although some non-relevant documents may also be retrieved (http://www.expasy.org/ch2d/2DHunt/).

So far five different 2D-PAGE databases focused on yeast organisms have been developed, namely: (i) SWISS-2DPAGE [53] from the Swiss Institute of Bioinformatics (Switzerland), (ii) Yeast Protein Map (YPM) [54,55] from the Institute of Cell Biochemistry and Genetic in Bordeaux (France), (iii) Yeast Proteome Database (YPD) [56–58] at Incyte Genomics, Inc. (California, US), (iv) YEAST 2D-PAGE at the University of Göteborg (Sweden) [59], and (v) COMPLUYEAST-2DPAGE at the Complutense University of Madrid (Spain). All of them contain reference 2-DE gel images from *Saccharomyces cerevisiae* (commonly known as baker's yeast), but the last one also includes annotated 2-DE maps from the human fungal pathogen *C. albicans* (see below).

In this section we highlight the noteworthy potential of the setting up of 2D-PAGE databases over the WWW network. Subsequently, we describe in detail the pioneering publicly federated 2D-PAGE database on *C. albicans* (namely, the COMPLUYEAST-2DPAGE database), remarking on its features and user interfaces as well as its current status and future challenges.

3.1. Reference 2-DE maps: the "virtual lab" of proteomics

The 2D-PAGE technology combined with protein identification allows the establishment of reference proteome maps. Additionally, the creation of these 2-DE maps is a powerful approach for (i) accurately defining which proteins are expressed in an organism or tissue under the influence of external or internal factors, (ii) revealing proteins exclusive to a subcellular compartment or tissue, (iii) mapping the gene products, whose open reading frames predicted in the genome sequence projects are an unknown function, to a specific location (i.e., as a complement to sequencing projects), (iv) suggesting potential or approximate 2-DE locations for proteins from other species boundaries with hitherto non-sequenced genomes (although it is important to take into account that isoelectric points of proteins are not often well conserved [60]), (v) monitoring gene expression levels or studying qualitative and quantitative changes in protein expression profiles, (vi) comparing modified protein patterns (in an attempt to establish which proteins are related with such changes), (vii) characterizing post-translational modifications under diverse expression conditions or in different strains, and (viii) defining putative interactions with other proteins, etc. [46].

In turn, these reference 2-DE maps (from one or several biological samples) can be stored, organized and displayed virtually in 2D-PAGE databases, enabling users to query and retrieve precise information pertaining to all identified 2-DE protein spots. Undoubtedly, the major underlying goal of these 2D-PAGE databases is to offer a direct and graphic way of linking genomes to proteomes [61]. In view of this, they could also be considered as a ''cyber-proteome-lab'', which provides scientists around the world working on similar biological material under the same specified physiological conditions with very important assets.

Accordingly, proteome researchers can remotely compare or match their experimental 2-DE gel patterns against the corresponding reference 2-DE maps present in such 2D-PAGE databases. This approach could permit users to (i) suggest putative identifications of any protein spot with relative ease, and/or (ii) investigate the presence of differences and similarities in the protein profiles between both 2-DE gel image patterns whenever these derive, of course, from the same type of biological specimen and sample preparation conditions. This can be achieved using different software packages. The free Flicker gel comparison program [62] (available on NCI web server at URL address http://www-lecb.ncifcrf.gov/flicker) and WebGel group-accessible software [63] (http://www-lecb.ncifcrf.gov/webgel) permit the analysis of 2-DE gels in the Intra- and Internet; more precisely, the comparison of two 2-DE maps retrieved from two different remote databases. Alternatively, other commercially

available 2-DE analysis software packages, such as Melanie, can automatically open a Web browser, query the remote 2D-PAGE database through the protein accession numbers, and recover specified information related to features on 2-DE-analyzed gels [64] (see below and Fig. 4).

3.2. COMPLUYEAST-2DPAGE database: the pioneering library of C. albicans subproteomes

To our knowledge, the COMPLUYEAST-2DPAGE database is currently the first and only publicly available 2D-PAGE library established for *C. albicans*. It is a databank created and locally maintained by our research team at the Complutense University's Microbiology Laboratory in Madrid (Spain). COMPLUYEAST-2DPAGE is freely available – in the same format as SWISSPROT-2DPAGE – on the ExPASy Web server, and can be accessed through the WWW network at its URL http://www.expasy.ch/ch2d/2d-index.html or http://babbage.csc.ucm.es/2d/2d.html.

3.2.1. Features and user interface. A federated 2D-PAGE database

The COMPLUYEAST-2DPAGE is a federated 2D-PAGE database that conforms to all the guidelines for standardization of the publication and sharing of its 2-DE data set over the Net. As reported in [47] (also available at http://www.expasy.ch/ch2d/fed-rules.html), the rules for the setting up of such databanks on a Web site are precisely (i) the inclusion of criteria of accessibility to each individual entry in the database by means of remote keyword search (rule 1), (ii) links with other databases (or at least with the main index) through active hypertext cross-references (rule 2), (iii) use of a main index (e.g., SWISS-PROT) by carrying cross-references to it (bidirectional linkages) (rule 3), (iv) the inclusion of further accessibility to each individual protein spot entry through clickable 2-DE images (rule 4), and (v) direct access from any 2-DE computer analysis software (designed for use with this type of databases) to individual entries (rule 5). Fig. 4 illustrates several COMPLUYEAST-2DPAGE Web pages that demonstrate the full implementation of these defined rules for building a federated 2D-PAGE database.

As stated above, there are two different ways to query this 2-DE data collection on *C. albicans*, i.e., (i) by a keyword search (rule 1) or (ii) by spot clicking (rule 4). Alternatively, a complete list of all COMPLUYEAST-2DPAGE entries in alphabetic order can also be displayed, enabling external users to recover the data on any protein of their interest with ease (Fig. 5).

In the first option, a particular protein can be searched for by means of different keywords, such as (i) the protein name, i.e., any word or partial word appearing in the description lines (ED) from the information available about the given protein (see below), (ii) the SWISS-PROT entry name, as reported in our papers (ID lines), (iii) the SWISS-PROT accession number (AC lines), and (iv) a referenced author name (RA lines). A query using any of these keywords will retrieve the corresponding searched protein identity Web page (Fig. 5).

Graphical query interfaces have been developed to implement the second approach. In so doing, interactive protein spots on the 2-DE maps are queried. This is achieved by clicking on a red-highlighted spot from the selected reference map. However, it is also possible to select the same map in another format (large or small format with highlighted or non-highlighted spots). The identified protein entry page present in this database is displayed immediately (Fig. 5). This interactive user interface therefore also exploits the concept ''click and tell''.

Each retrieved protein entry page provides (i) data on its subcellular location (e.g., cytoplasmic extracts, protoplast lysates and proteins secreted by regenerating protoplasts), its bibliographical references and its experimental pI and M_w values, and (ii) a cross-reference to its specific entry in the SWISS-PROT database, which in turn is linked to several other public data collections – see Section 2.1.1 – thereby providing detailed information about the protein in question (rule 2 and 3), and (iii) the 2-DE map(s) that include(s) the same protein (protein mapping). By clicking on any available 2-DE image, it is possible to visualize the location, highlighted in red, of this known protein on the 2-DE map – or on the different subproteome maps if there are several

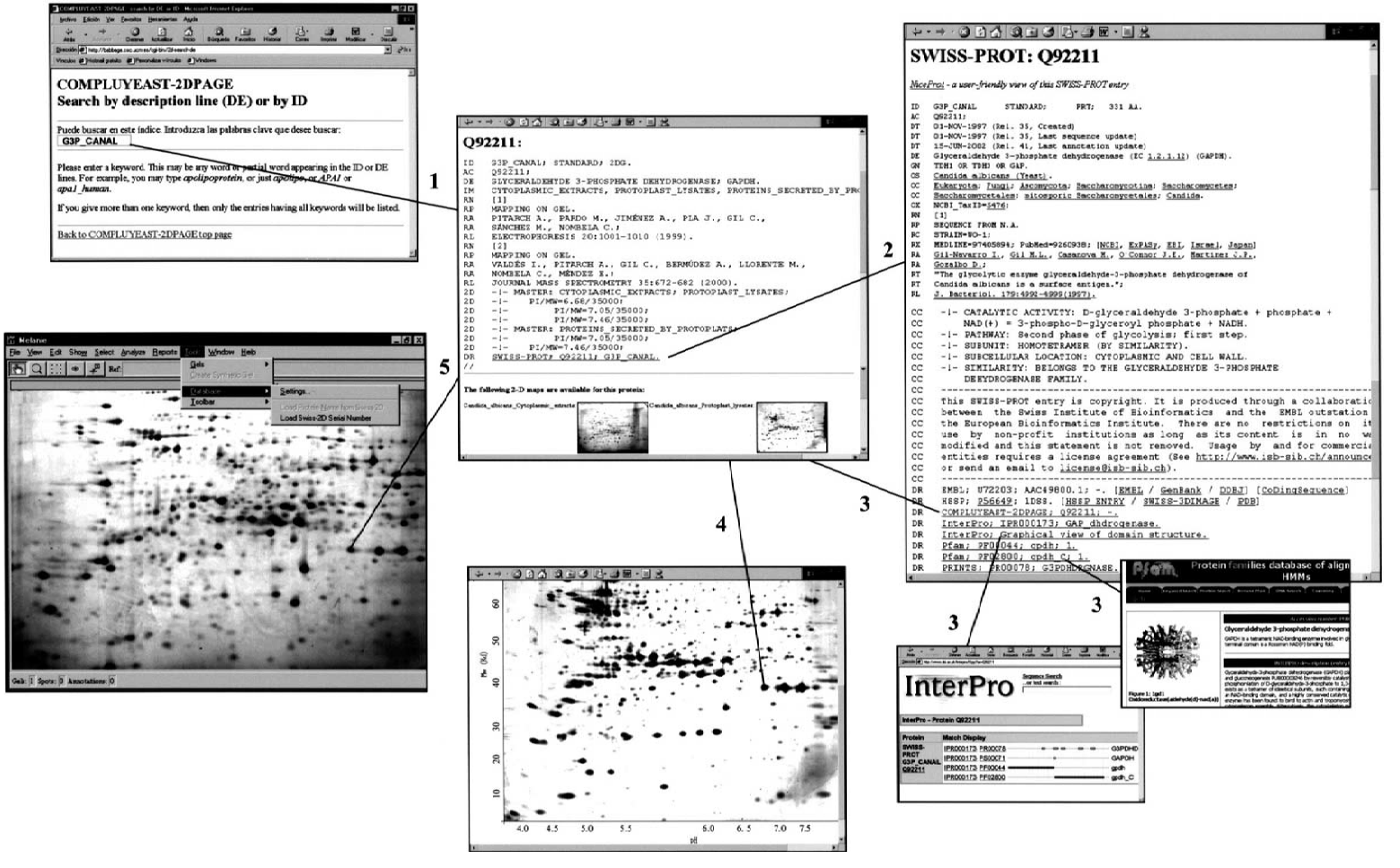

Fig. 4. COMPLUYEAST-2DPAGE's web-browser interface. This 2D-PAGE data collection is an example of a federated database. Numbers correspond to each of the five rules for the building of such databanks (see text).

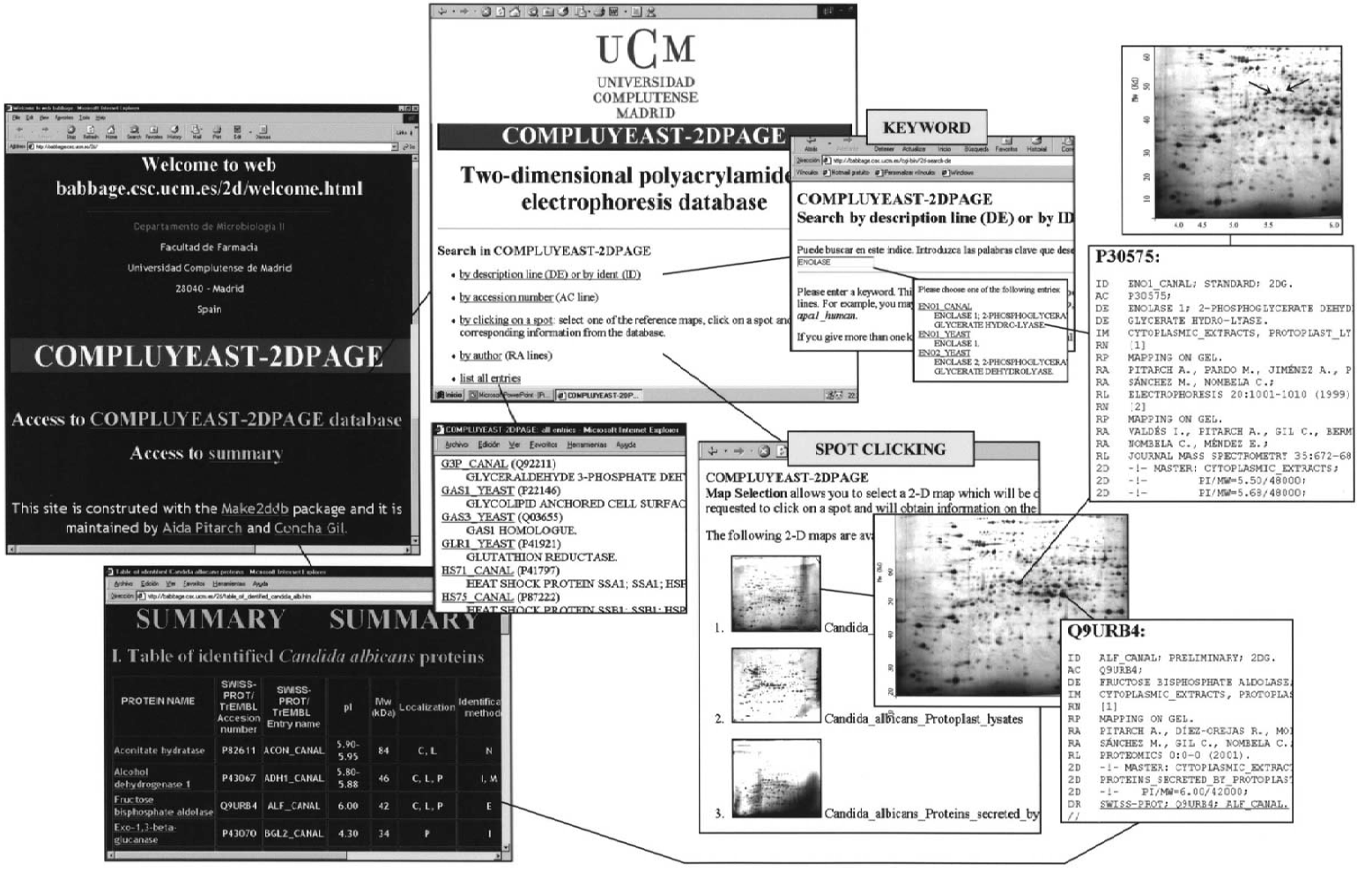

Fig. 5. Different ways to access any given entry present in the COMPLUYEAST-2DPAGE database.

Table 2
Summary of identified *C. albicans* proteins present in COMPLUYEAST-2DPAGE database

Protein name	SWISS-Prot/TrEMBL Accession number	SWISS-Prot/TrEMBL Entry name	p*I*	M_w (kDa)	Subcellular location*	Identification method(s)	References
Aconitate hydratase	P82611	ACON_CANAL	5.90–5.95	84	C, L	MS/MS	[22]
Alcohol dehydrogenase 1	P43067	ADH1_CANAL	5.80–5.88	46	C, L, P	Immunoblotting MALDI MS	[35,67]
Fructose bisphosphate aldolase	Q9URB4	ALF_CANAL	6.00	42	C, L, P	Edman sequencing	[35]
Exo-1,3-beta-glucanase	P43070	BGL2_CANAL	4.30	34	P	Immunoblotting	[65]
Enolase1	P30575	ENO1_CANAL	5.31–5.76	48	C, L, P	Immunoblotting MALDI MS	[65,67]
Glyceraldehyde-3-phosphate dehydrogenase	Q92211	G3P_CANAL	6.68–7.46	35	C, L, P	Immunoblotting MALDI MS	[65,67]
Heat shock protein SSA1	P41797	HS71_CANAL	5.06–5.17	71	C, L	MALDI MS	[67]
Heat shock protein SSB1	P87222	HS75_CANAL	4.87–5.14	67	C, L	MALDI MS	[67]
Inosine-5′-monophosphate dehydrogenase	O00086	IMH3_CANAL	6.70	56	C, L, P	Edman sequencing	[35]
Pyruvate kinase	P46614	KPYK_CANAL	7.00	58	C, L, P	MS/MS	[22]
Methionine synthase	P82610	METE_CANAL	5.38–5.70	84	C, L, P	Edman sequencing MS/MS	[22,35]
Phosphoglycerate kinase	P46273	PGK_CANAL	5.76–6.16	46	C, L, P	Immunoblotting Edman sequencing MALDI MS	[22,35,67]
Phosphoglycerate mutase 1	P82612	PMG1_CANAL	5.90	29	C	Edman sequencing MS/MS	[22]
Phosphomannomutase	P31353	PMM_CANAL	5.24	29	C, L	MALDI MS	[67]
Triosephosphate isomerase	P82613	TPIS_CANAL	5.89	30	C, L	Edman degradation	[35]

*C, L and P refer to cytoplasmic extracts, protoplast lysates and proteins secreted by protoplasts.

images (Fig. 5). All *C. albicans* COMPLUYEAST-2DPAGE entries have been designed in an accessible uniform format.

The first Web page from this database also grants access to a summary that encloses a table of all *C. albicans* proteins (from different subproteomes) identified by several different methods. The SWISS-PROT/TrEMBL accession number and entry name, the experimental isoelectric point and molecular weight (deduced from 2-DE gels using Melanie analysis software system), the subcellular location (i.e., cell wall or cytoplasmic compartment), and method of identification of these proteins are listed on this Web page (see Fig. 5) and are also shown in Table 2.

This 2D-PAGE database currently contains three *C. albicans* reference 2-DE maps pertaining to cytoplasmic extracts, protoplast lysates and proteins

secreted by protoplasts when they regenerate their walls. Non-linear wide pH gradient (3–10 NL IPG) strips were used as first-dimension separation [65]. Fifteen protein entries for *C. albicans* are available at the COMPLUYEAST-2DPAGE, and have been characterized in terms of their p*I* and M_r values and by means of (i) immunoblotting (5 spots), (ii) N-terminal microsequencing or Edman degradation (6 spots), (iii) MALDI-TOF (matrix-assisted laser desorption/ionization time-of-flight) mass spectrometry–mass peptide fingerprinting (7 spots), and (iv) nanoelectrospray–tandem mass spectrometry (nanoES-MS/MS) – peptide sequences (4 spots).

Further characterization of the proteins from the above cellular compartments and other specific subproteomes (different cell wall fractions from *C. albicans* yeast and hyphal forms [66]) is currently being addressed at our laboratory using mass spectrometry. These identifications will be deposited in this database as soon as possible. Due to the forthcoming completion of annotated *C. albicans* sequence databases, a higher number of protein identifications is expected to be achieved in the foreseeable future.

3.2.2. An outlook

This 2-DE database adds a new dimension to the study of *C. albicans* genome expression. A previous edition of the maps and characterization of protein spots present in COMPLUYEAST-2DPAGE has been published in several international journals [22,35,65,67] (Table 2).

Owing to the nature of the identified proteins, COMPLUYEAST-2D-PAGE offers an earlier data collection on proteins associated with (i) antigenic properties (i.e., some of these proteins have been detected using sera from infected mice or patients with systemic candidiasis [22,35,65]), (ii) metabolic and secretory pathways (i.e., some proteins are included in the glycolytic pathway, while others are secreted into the medium during the regeneration of protoplast cell walls [22,35,65,67]), and (iii) cell wall and/or cellular surface [65], to name but a few. It is thus our hope that this data set could offer a useful contribution to the scientific community working on the *C. albicans* proteome or on some of the reported issues. For instance, the available reference 2-DE maps may be powerful tools for the study of the glycolytic pathway or cell wall mutants, providing information on which proteins are altered as result of a particular gene deletion by means of analysis of the presence, absence and/or synthesis levels of proteins in both mutant and wild-type patterns. However our 2D-PAGE database will be regularly updated as new protein identifications or subproteome maps become available.

4. The "cyber-bioknowledge library" of the complex world of proteins. New frontiers and challenges in the third millennium

The rapidly evolving field of bioinformatics and artificial intelligence heralds the advent of innovative breakthroughs in the understanding of biological events in cells and organisms and, in turn, in the development of new therapeutic strategies. In so doing, scientists draw on bioinformatics to manage the huge and complex data sets derived from both genome sequencing and proteomic projects at large scale. Accordingly, information science and electronic data are currently replacing the basically lab-based science. In parallel, in an era rich in biological data, a new means of worldwide communication – the World Wide Web – has certainly facilitated the integration and diffusion, whenever required, of a huge assortment of databases containing this deluge of information among the scientific community.

Increasingly, the genomes of many organisms are being fully sequenced and in the near future researchers should be able to witness a comprehensive characterization of their proteomes. Proteomics, together with the assistance of bioinformatics, could offer new insight into the complexity of biological systems. Also, the completion of the genome sequences of different organisms should enable the simultaneous study of thousands of genes using transcript profilings (transcriptomes). In this way, for instance, the Galar Fungail European Network is currently developing *C. albicans* DNA chips in an attempt to analyze fungal virulence factors and host–fungus interactions [68].

To gain meaningful clues into the basic networks that organize and drive the life of cells and organisms, the design of a broad variety of sophisticated bioinformatic tools will undoubtedly become neces-

sary, although some are already available (e.g., in the ExPASy server), which translate all raw data generated into knowledge. Such tools could therefore allow expert users to (i) analyze and interpret all the data generated, (ii) compare expression profiles, (iii) correlate proteomes with transcriptomes, (iv) elucidate post-translational modifications in proteins, (v) deduce potential functions, (vi) suggest systems of regulation, (vii) predict the three-dimensional structure of each protein – from data derived from X-ray crystallography, nuclear magnetic resonance and electron microscopy studies – (creation of 3-D molecular models), (viii) correlate sequence information with protein structure and function, (ix) link proteomes to metabolism (the metabolome) and signaling pathways (the signalome) and (x) establish interactions among proteins in an effort to model biological processes within and among cells (the interactome [69]) (Fig. 1). In view of this, these novel approaches could play a key role in functional knowledge about the complex world of proteins at different levels; i.e., of the biomolecules involved in nearly all processes occurring in living cells.

New specialized types of databases will also become necessary to cater to all this information. Some pioneering databanks have already been developed and are available through the Web. Nevertheless, there is no doubt that a correct combination of these data sets could afford new points of view concerning the identification of genes responsible for diseases and pathogenic mechanisms, and could facilitate the discovery of new diagnostic markers, drug targets, and therapies. As a result, databases currently constitute the most important and fruitful approach for the understanding of life processes. The paramount challenge in bioinformatics is thus to combine all these databases over the Internet. Such a merger will culminate in the foremost virtual biological lab or ‘‘cyber-bioknowledge library’’ that life-science researchers will be able to exploit in years to come. In this way, the future of life science essentially depends on the degree and coverage of database integration.

So far, we only know which pieces make up the puzzle of the life, but the assembling of all these pieces of the puzzle continues to be a challenge for biological scientists and indeed there is still much hard work to be done.

5. Nomenclature

HTML	Hyper Text Markup Language
URL	Uniform Resource Locator
WWW	World Wide Web
FTP	File Transfer Protocol
2D-PAGE	two-dimensional polyacrylamide gel electrophoresis
2-DE	two-dimensional electrophoresis
MS	mass spectrometry
SRS	Sequence Retrieval System
ExPASy	Expert Protein Analysis System

Acknowledgements

We thank Dr. LaJean Chaffin for critical reading of the manuscript. This work was supported by grants: SAF 2000-0108 from Comisión Interministerial de Ciencia y Tecnología (CICYT), CPGE 1010/2000 Strategic Groups from Comunidad Autónoma de Madrid, and 12th National Contest on ‘‘Functional Genomics, Proteomics and Pharmacogenetics’’ from Fundación Ramón Areces (Spain). A. Pitarch was the recipient of a Fellowship from Fundación Ramón Areces of Spain.

References

[1] J.F. Ernst, Microbiology 146 (2000) 1763.
[2] N.A. Gow, Curr. Top. Med. Mycol. 8 (1997) 43.
[3] G. Garber, Drugs 61 (2001) 1.
[4] P. Sandven, Rev. Iberoam Micol. 17 (2000) 73.
[5] J.L. Vincent, E. Anaissie, H. Bruining, W. Demajo, M. el Ebiary, J. Haber, Y. Hiramatsu, G. Nitenberg, P.O. Nystrom, D. Pittet, T. Rogers, P. Sandven, G. Sganga, M.D. Schaller, J. Solomkin, Intensive Care Med. 24 (1998) 206.
[6] D. Sanglard, F.C. Odds, Lancet Infect. Dis. 2 (2002) 73.
[7] J. Pla, C. Gil, L. Monteoliva, F. Navarro-Garcia, M. Sanchez, C. Nombela, Yeast 12 (1996) 1677.
[8] M.D. De Backer, P.T. Magee, J. Pla, Annu. Rev. Microbiol. 54 (2000) 463.
[9] M. Niimi, R.D. Cannon, B.C. Monk, Electrophoresis 20 (1999) 2299.
[10] A. Pitarch, M. Sánchez, C. Nombela, C. Gil, J. Chromatogr. B 787 (2003) 101.
[11] R.D. Fleischmann, M.D. Adams, O. White, R.A. Clayton, E.F. Kirkness, A.R. Kerlavage, C.J. Bult, J.F. Tomb, B.A. Dougherty, J.M. Merrick, Science 269 (1995) 496.
[12] E.S. Lander et al., Nature 409 (2001) 860.

[13] J.C. Venter et al., Science 291 (2001) 1304.
[14] A. Goffeau, B.G. Barrell, H. Bussey, R.W. Davis, B. Dujon, H. Feldmann, F. Galibert, J.D. Hoheisel, C. Jacq, M. Johnston, E.J. Louis, H.W. Mewes, Y. Murakami, P. Philippsen, H. Tettelin, S.G. Oliver, Science 274 (1996) 546.
[15] V. Wood, R.G. William, M.A. Rajandream, M. Lyne, R. Lyne, A. Stewart, J. Sgouros, N. Peat, J. Hayles, S. Baker et al., Nature 415 (2002) 871.
[16] A.M. Gillum, E.Y. Tsay, D.R. Kirsch, Mol. Gen. Genet. 198 (1984) 179.
[17] A.K. Goshorn, S. Scherer, Genetics 123 (1989) 667.
[18] E. Tait, M.C. Simon, S. King, A.J. Brown, N.A. Gow, D.J. Shaw, Fungal Genet. Biol. 21 (1997) 308.
[19] H. Chibana, B.B. Magee, S. Grindle, Y. Ran, S. Scherer, P.T. Magee, Genetics 149 (1998) 1739.
[20] M.R. Wilkins, C. Pasquali, R.D. Appel, K. Ou, O. Golaz, J.C. Sánchez, J.X. Yan, A.A. Gooley, G.J. Hughes, I. Humphery-Smith, K.L. Williams, D.F. Hochstrasser, Biotechnology 14 (1996) 61.
[21] P.R. Graves, T.A. Haystead, Microbiol. Mol. Biol. Rev. 66 (2002) 39.
[22] M. Pardo, M. Ward, A. Pitarch, M. Sánchez, C. Nombela, W. Blackstock, C. Gil, Electrophoresis 21 (2000) 2651.
[23] Chapter 6 R.D. Appel, in: M.R. Wilkins, K.L. Williams, R.D. Appel, D.F. Hochstrasser (Eds.), Proteome Research: New Frontiers in Functional Genomics, Springer-Verlag, Berlin, 1997, p. 149.
[24] R.D. Appel, A. Bairoch, D.F. Hochstrasser, Trends Biochem. Sci. 19 (1994) 258.
[25] H. Recipon, W. Makalowski, Curr. Opin. Biotechnol. 8 (1997) 115.
[26] Chapter 5 A. Bairoch, in: M.R. Wilkins, K.L. Williams, R.D. Appel, D.F. Hochstrasser (Eds.), Proteome Research: New Frontiers in Functional Genomics, Springer-Verlag, Berlin, 1997, p. 93.
[27] J.D. Cavalcoli, Trends Cardiovasc. Med. 11 (2001) 76.
[28] G.D. Schuler, J.A. Epstein, H. Ohkawa, J.A. Kans, Methods Enzymol. 266 (1996) 141.
[29] T. Etzold, A. Ulyanov, P. Argos, Methods Enzymol. 266 (1996) 114.
[30] A. Bairoch, R. Apweiler, Nucleic Acids Res. 28 (2000) 45.
[31] E. Gasteiger, E. Jung, A. Bairoch, Curr. Issues Mol. Biol. 3 (2001) 47.
[32] A. Bairoch, R. Apweiler, Nucleic Acids Res. 25 (1997) 31.
[33] R. Apweiler, A. Gateau, S. Contrino, M.J. Martin, V. Junker, C. O'Donovan, F. Lang, N. Mitaritonna, S. Kappus, A. Bairoch, Proc. Int. Conf. Intell. Syst. Mol. Biol. 5 (1997) 33.
[34] V. Junker, S. Contrino, W. Fleischmann, H. Hermjakob, F. Lang, M. Magrane, M.J. Martin, N. Mitaritonna, C. O'Donovan, R. Apweiler, J. Biotechnol. 78 (2000) 221.
[35] A. Pitarch, R. Díez-Orejas, G. Molero, M. Pardo, M. Sánchez, C. Gil, C. Nombela, Proteomics 1 (2001) 550.
[36] C.H. Wu, H. Huang, L. Arminski, J. Castro-Alvear, Y. Chen, Z.Z. Hu, R.S. Ledley, K.C. Lewis, H.W. Mewes, B.C. Orcutt, B.E. Suzek, A. Tsugita, C.R. Vinayaka, L.S. Yeh, J. Zhang, W.C. Barker, Nucleic Acids Res. 30 (2002) 35.
[37] A.J. Bleasby, D. Akrigg, T.K. Attwood, Nucleic Acids Res. 22 (1994) 3574.
[38] S.F. Altschul, W. Gish, W. Miller, E.W. Myers, D.J. Lipman, J. Mol. Biol. 215 (1990) 403.
[39] M.C. Costanzo, J.D. Hogan, M.E. Cusick, B.P. Davis, A.M. Fancher, P.E. Hodges, P. Kondu, C. Lengieza, J.E. Lew-Smith, C. Lingner, K.J. Roberg-Perez, M. Tillberg, J.E. Brooks, J.I. Garrels, Nucleic Acids Res. 28 (2000) 73.
[40] M.C. Costanzo, M.E. Crawford, J.E. Hirschman, J.E. Kranz, P. Olsen, L.S. Robertson, M.S. Skrzypek, B.R. Braun, K.L. Hopkins, P. Kondu, C. Lengieza, J.E. Lew-Smith, M. Tillberg, J.I. Garrels, Nucleic Acids Res. 29 (2001) 75.
[41] C. Csank, M.C. Costanzo, J. Hirschman, P. Hodges, J.E. Kranz, M. Mangan, K. O'Neill, L.S. Robertson, M.S. Skrzypek, J. Brooks, J.I. Garrels, Methods Enzymol. 350 (2002) 347.
[42] P.E. Hodges, W.E. Payne, J.I. Garrels, Nucleic Acids Res. 26 (1998) 68.
[43] J.E. Celis, P. Gromov, Curr. Opin. Biotechnol. 10 (1999) 16.
[44] R.D. Appel, D.F. Hochstrasser, M. Funk, J.R. Vargas, C. Pellegrini, A.F. Muller, J.R. Scherrer, Electrophoresis 12 (1991) 722.
[45] J.I. Garrels, J. Biol. Chem. 264 (1989) 5269.
[46] Chapter 4 M.R. Wilkins, A.A. Gooley, in: M.R. Wilkins, K.L. Williams, R.D. Appel, D.F. Hochstrasser (Eds.), Proteome Research: New Frontiers in Functional Genomics, Springer-Verlag, Berlin, 1997, p. 65.
[47] R.D. Appel, A. Bairoch, J.C. Sanchez, J.R. Vargas, O. Golaz, C. Pasquali, D.F. Hochstrasser, Electrophoresis 17 (1996) 540.
[48] R.D. Appel, J.C. Sanchez, A. Bairoch, O. Golaz, M. Miu, J.R. Vargas, D.F. Hochstrasser, Electrophoresis 14 (1993) 1232.
[49] J.C. Sanchez, R.D. Appel, O. Golaz, C. Pasquali, F. Ravier, A. Bairoch, D.F. Hochstrasser, Electrophoresis 16 (1995) 1131.
[50] C. Hoogland, J.C. Sanchez, L. Tonella, P.A. Binz, A. Bairoch, D.F. Hochstrasser, R.D. Appel, Nucleic Acids Res. 28 (2000) 286.
[51] R.D. Appel, A. Bairoch, D.F. Hochstrasser, Methods Mol. Biol. 112 (1999) 383.
[52] M. Vihinen, Bioinformatics in proteomics, Biomol. Eng. 18 (2001) 241.
[53] J.C. Sanchez, O. Golaz, S. Frutiger, D. Schaller, R.D. Appel, A. Bairoch, G.J. Hughes, D.F. Hochstrasser, Electrophoresis 17 (1996) 556.
[54] H. Boucherie, F. Sagliocco, R. Joubert, I. Maillet, J. Labarre, M. Perrot, Electrophoresis 17 (1996) 1683.
[55] M. Perrot, F. Sagliocco, T. Mini, C. Monribot, U. Schneider, A. Shevchenko, M. Mann, P. Jeno, H. Boucherie, Electrophoresis 20 (1999) 2280.
[56] W.E. Payne, J.I. Garrels, Nucleic Acids Res. 25 (1997) 57.
[57] J.I. Garrels, C.S. McLaughlin, J.R. Warner, B. Futcher, G.I. Latter, R. Kobayashi, B. Schwender, T. Volpe, D.S. Anderson, R. Mesquita-Fuentes, W.E. Payne, Electrophoresis 18 (1997) 1347.
[58] P.E. Hodges, A.H. McKee, B.P. Davis, W.E. Payne, J.I. Garrels, Nucleic Acids Res. 27 (1999) 69.
[59] J. Norbeck, A. Blomberg, Yeast 13 (1997) 1519.

[60] M.R. Wilkins, K.L. Williams, J. Theor. Biol. 186 (1997) 7.
[61] M.R. Wilkins, D.F. Hochstrasser, J.C. Sanchez, A. Bairoch, R.D. Appel, Trends Biochem. Sci. 21 (1996) 496.
[62] P.F. Lemkin, Electrophoresis 18 (1997) 461.
[63] P.F. Lemkin, J.M. Myrick, Y. Lakshmanan, M.J. Shue, J.L. Patrick, P.V. Hornbeck, G.C. Thornwal, A.W. Partin, Electrophoresis 20 (1999) 3492.
[64] R.D. Appel, P.M. Palagi, D. Walther, J.R. Vargas, J.C. Sanchez, F. Ravier, C. Pasquali, D.F. Hochstrasser, Electrophoresis 18 (1997) 2724.
[65] A. Pitarch, M. Pardo, A. Jimenez, J. Pla, C. Gil, M. Sánchez, C. Nombela, Electrophoresis 20 (1999) 1001.
[66] A. Pitarch, M. Sánchez, C. Nombela, C. Gil, submitted for publication.
[67] I. Valdes, A. Pitarch, C. Gil, A. Bermudez, M. Llorente, C. Nombela, E. Mendez, J. Mass Spectrom. 35 (2000) 672.
[68] A.M. Murad, C. d'Enfert, C. Gaillardin, H. Tournu, F. Tekaia, D. Talibi, D. Marechal, V. Marchais, J. Cottin, A.J. Brown, Mol. Microbiol. 42 (2001) 981.
[69] M. Gerstein, N. Lan, R. Jansen, Science 295 (2002) 284.