The Power and the Limitations of Cross-Species Protein Identification by Mass Spectrometry-driven Sequence Similarity Searches*[S]

Bianca Habermann‡§¶, Jeffrey Oegema§, Shamil Sunyaev‖, and Andrej Shevchenko‡¶

Mass spectrometry-driven BLAST (MS BLAST) is a database search protocol for identifying unknown proteins by sequence similarity to homologous proteins available in a database. MS BLAST utilizes redundant, degenerate, and partially inaccurate peptide sequence data obtained by *de novo* interpretation of tandem mass spectra and has become a powerful tool in functional proteomic research. Using computational modeling, we evaluated the potential of MS BLAST for proteome-wide identification of unknown proteins. We determined how the success rate of protein identification depends on the full-length sequence identity between the queried protein and its closest homologue in a database. We also estimated phylogenetic distances between organisms under study and related reference organisms with completely sequenced genomes that allow substantial coverage of unknown proteomes. ***Molecular & Cellular Proteomics 3:238–249, 2004.***

Proteomics has become a powerful tool to understand the function and regulation of genes through the large-scale study of proteins in living cells (reviewed in Refs. 1–4). Proteomics efforts are supported by the identification of proteins and their post-translational modifications by mass spectrometry, as it offers the femtomole sensitivity, high throughput, and is able to decipher complex mixtures of proteins. Proteins are typically digested in-gel or in-solution with proteolytic enzymes, and the digests are analyzed by peptide mass mapping and/or tandem mass spectrometry (reviewed in Refs. 4 and 5). Conventional methods of database searching heavily rely on matching masses of intact peptides (peptide mass mapping) or their fragments (tandem mass spectrometry) to the corresponding masses of peptides and/or peptide fragments obtained by *in silico* processing of protein sequences from database entries (reviewed in Ref. 6). Stringent matching of computed and measured masses dramatically increases the specificity and the speed of database searching (7), yet restricts the reach of proteomics down to a handful of model species, for which either a complete genome and/or a substantial number of cDNA sequences is available in a database. Despite spectacular progress of genomic sequencing, many important model organisms yet have not been adequately covered (8).

If a protein of interest is not present in a database, peptide sequences can be deduced by *de novo* interpretation of tandem mass spectra (reviewed in Ref. 9) and used for designing degenerate oligonulcleotide probes. The cognate gene can subsequently be cloned by a PCR-based method. However, cloning experiments are expensive, laborious, require long and accurate stretches of peptide sequence, and, despite previously demonstrated success (10–13), have never been applied for the high-throughput characterization of proteomes.

Peptide sequences can also be employed in identifying proteins by sequence similarity searches (14–17). These search methods represent an attractive alternative to cloning because the identification of unknown proteins can be achieved without further "wet" biochemistry experiments, and it is possible to utilize less-accurately determined peptide sequences (reviewed in Refs. 8 and 18). However, mass spectrometry and sequence similarity searches are difficult to combine. Conventional database search algorithms like BLAST (19) or FASTA (20) are optimized for accurate sequence queries that are longer than 35 amino acid residues (21, 22). Usually peptide sequences obtained by tandem mass spectrometry do not exceed the length of a tryptic peptide, typically comprising 10–15 amino acid residues, and therefore the statistical significance of retrieved hits is often ambiguous.

Recently, several database searching approaches were reported that accommodate specific requirements of tandem mass spectrometric sequencing (14–17, 23). Shevchenko *et al.* developed a BLAST2-based search protocol termed MS BLAST (15). MS BLAST takes advantage of several search options in WU-BLAST2 (21, 24) and employs a scoring matrix optimized for peptide sequences produced by tandem mass spectrometry. MS BLAST does not allow gaps within individual peptides, while gaps between peptides are not penalized

From the ‡Max Planck Institute of Molecular Cell Biology and Genetics, Pfotenhauerstrasse 108, 01307 Dresden, Germany; and ‖Birgham & Women's Hospital and Harvard Medical School, 75 Francis Street, Boston, MA 02115

Received, July 31, 2003, and in revised form, December 22, 2003
Published, MCP Papers in Press, December 26, 2003, DOI 10.1074/mcp.M300073-MCP200

and can be of arbitrary length. Therefore all peptide sequences obtained by the interpretation of acquired tandem mass spectra are assembled into a single searching string in arbitrary order (18). MS BLAST identifies a set of high-scoring segment pairs (HSPs)[1] between the queried peptides and sequences from database entries and scores these HSPs independently of their respective location on a protein backbone and in the queried string. Because the smallest sum probability computed by WU-BLAST2 does not adequately merit the statistical significance of reported hits, the MS BLAST scoring scheme maximizes the raw score, rather than minimizing the smallest sum probability of reported alignments. The total score of the hit is additive over scores of individual HSPs and is then compared with the precomputed significance thresholds. The *span1* filter replaces multiple HSPs aligned to the same segment of a database sequence by a single HSP with the highest score. Therefore, an MS BLAST query can contain hundreds of redundant, degenerate, and partially accurate peptide sequence candidates and can directly import the output of automated *de novo* interpretation of multiple tandem mass spectra obtained in liquid chromatography tandem mass spectrometry (LC MS/MS) (25, 26), nanoelectrospray tandem mass spectrometry (Nano ES MS/MS) (18), or matrix-assisted laser desorption/ionization tandem mass spectrometry (MALDI-MS/MS) (27, 28) experiments.

To streamline the analysis of the output of MS BLAST database searches, a parsing script was developed to evaluate and sort hits according to the MS BLAST scoring scheme. Regardless of their total scores (which depends on the number of aligned HSPs), hits significant in MS BLAST sense are color-coded and placed at the top of the output list. MS BLAST running with this parsing script has been installed at a web-accessible server (dove.embl-heidelberg.de/Blast2/msblast.html).

Modified FASTA-based algorithms, such as FASTS and FASTF (14, 16, 17), evaluate hits by original scoring procedures and statistical significance criteria. Despite higher flexibility (allowing gapped and nongapped alignments, consideration of isobaric permutations in peptide sequences, and other useful features), FASTA-based search software requires time-intense computations, and the significance of hits declines with increasing numbers of redundant peptide sequence candidates in the query.

BLAST and FASTA-based approaches have been successfully applied to identify proteins from organisms with unsequenced genomes using peptide queries generated by mass spectrometry. Comparative testing of MS-Shotgun, FASTS, and MS BLAST on a small dataset of peptide sequences from 14 proteins of the 20S proteasome of *Trypanosoma brucei* suggested similar performance of these three search engines (16, 17).[2] In a recent study, sequence similarity searches by MS BLAST almost doubled the number of identified microtubule-associated proteins from African clawed frog *Xenopus laevis* (29) compared with a conventional database searching method that utilizes stringent cross-species matching of uninterpreted tandem mass spectra to peptides from database entries (30). However, no evidence is yet available if sequence similarity identification methods might have a significant impact on the characterization of entire proteomes. It is not clear what percentage of sequence identity to homologous proteins in a database is required for the identification of yet unknown proteins. In a broader perspective, it is not known what phylogenetic distance between a studied organism and reference organism(s) with sequenced genomes enables substantial coverage of its proteome. It is equally difficult to estimate what length and number of fragmented peptides would be sufficient for identifying homologous proteins by mass spectrometry-driven sequence database searches and how accurate *de novo* sequencing should be. We applied computational modeling to evaluate the potential of the MS BLAST protocol for the cross-species identification of proteins. We estimated how the success rate of protein identification depends on the full-length sequence identity between the queried protein and its closest homologue in a database. By evaluating the success rate of protein identification on the proteome scale, we estimated acceptable phylogenetic distances between an organism under study and related reference organisms with completely sequenced genomes.

EXPERIMENTAL PROCEDURES

Computer Simulation Experiments—The WU-BLAST2 program (24) was installed on a local server. Three species were selected from the fungal (*Sacharomyces cerevisiae*, *Candida albicans*, and *Schizosaccharomyces pombe*) and vertebrate (*Takifugu rubripes*, *Mus musculus*, and *Homo sapiens*) lineages. Full-length WU-BLAST2 searches and MS BLAST searches were carried out between the members of each lineage, so that proteins from each of the species were searched against the protein databases of the other two species in the same lineage (see Fig. 1*A*). One thousand proteins from *S. cerevisiae* from chromosomes II, X, and XIV, 1,000 proteins from *C. albicans*, as well as 1,000 proteins from *S. pombe* were randomly selected for the fungal group (see Fig. 1*B*). Five hundred proteins each from *T. rubripes*, *M. musculus*, and *H. sapiens* were randomly selected for the vertebrate group. Low-complexity regions from protein queries were filtered with *pseg* (31). Homologues of queried proteins in the neighboring proteomes were determined by WU-BLAST2 searches performed under standard settings (substitution matrix BLOSUM62, Expect cutoff 1) (21, 24) using their full-length sequences, and hits with E-values lower than 1E-05 were fetched from the output by a special sorting script. Sequence identity between the queried protein and retrieved hits was expressed as the percentage of identical residues normalized to the length of the query. To simulate MS BLAST queries, peptide sequences of 10 amino acid residues were randomly selected from proteins and merged into search strings. Queries containing 1, 3,

[1] The abbreviations used are: HSP, high-scoring segment pair; LC MS/MS, liquid chromatography tandem mass spectrometry; Nano ES MS/MS, nanoelectrospray tandem mass spectrometry.

[2] B. Habermann, S. Sunyaev, and A. Shevchenko, unpublished observations.

5, 8, 10, 15, and 20 unique peptides were assembled from peptide sequences from *S. cerevisiae* and *C. albicans*, and queries containing 3, 8, and 15 unique peptides were assembled from *S. pombe* proteins and from the three vertebrate species. To simulate possible ambiguities of *de novo* interpretation of tandem mass spectra, one or two randomly selected amino acid residues in each peptide sequence were replaced with an X symbol, which has a score of 0 in the PAM30MS substitution matrix. MS BLAST searches with assembled queries were performed as described previously (15) with the exception that the Expect cutoff was 1,000. In order to avoid a bias resulting from random selection of peptides for MS BLAST queries, peptide selection for each protein sequence was repeated five times, resulting in 5,000 MS BLAST queries for the fungal species and 2,500 MS BLAST queries for the vertebrates.

Calculation of Threshold Scores of Statistical Significance for the MS BLAST Scoring Matrix—To determine the thresholds of statistical significance of MS BLAST hits, we analyzed raw scores of nonrelated peptide sequence alignments essentially as described previously (15). Thresholds were calculated by performing 5,000 MS BLAST searches for each query composed from a given number of peptide sequences. The number of peptides in queries was within the range from 1 to 20. Queries were assembled from 10 amino acid residues peptides, which were obtained by five independent rounds of random selection from 1,000 unique proteins. MS BLAST queries were searched against an inverted comprehensive nonredundant database. The source database (release February, 2003), comprising 1,339,046 entries (644,844,000 amino acid residues), was downloaded from the National Center for Biotechnology Information. Scores of top hits were collected in a Microsoft Excel spreadsheet and sorted by the number of peptides in the query and by the number of reported HSPs. For each size of the query (ranging from 1 to 20 peptides), threshold scores were determined so that they exceeded scores of best hits (with a given number of HSPs) of 99% of searches. The table of precomputed threshold scores is available in the supplemental materials (Table 1S).

Threshold scores control the rate of expected false-positive hits, but not the rate of false-negative hits, and are independent of the composition of search queries. Calculating thresholds from searches against an inverted comprehensive nonredundant database and employing them to evaluate searches against much smaller species-specific databases provided a conserved estimate of MS BLAST performance. The large sample size of a nonredundant database also represented "averaged" statistical properties of many known proteomes.

Evaluation of the Sensitivity and Specificity of MS BLAST Searches—The significance of hits was evaluated according to the MS BLAST scoring scheme as described previously (15, 18): for every reported hit, the score of the top-ranked HSP was compared with the corresponding threshold score for a single-matched HSP from the MS BLAST scoring table. If the score exceeded the threshold, the hit was considered positively identified. If the score was below the threshold, the score from the first- and second-ranked HSP were summed up. In case the summed score exceeded the threshold for two matched HSPs, the identification was positive. Otherwise, adding the third-ranking HSP and so forth continued the procedure. Examples of the application of the MS BLAST scoring scheme are provided in Table I. The dataset for the organisms *M. musculus* and *S. pombe* can be downloaded from www.mpi-cbg.de/~habermann. The complete dataset is available upon request.

Estimation of Evolutionary Distances—For evaluating evolutionary distances, a phylogenetic tree was constructed for the fungal and vertebrate lineages based on the sequence of the mitochondrial small-subunit ribosomal RNA. Multiple sequence alignments were constructed using the ClustalX program (32). The evolutionary distance between species was calculated using the program dnadist from the Phylip package (33).

RESULTS AND DISCUSSION

Sensitivity of MS BLAST Identification—We were interested in the MS BLAST performance in cross-species protein identification with peptide queries produced by the interpretation of tandem mass spectra (MS queries). The success rate of protein identification by sequence-similarity searches depends on the molecular properties of analyzed proteins, the evolutionary conservation between analyzed proteins and their homologues in a database, and on the employed analytical methodology (34). Many of these factors are poorly understood and could not be controlled directly while constructing a dataset. To create a dataset that adequately mimics MS queries, we first generated a set of protein sequences that adequately represents the entire proteome of a model organism. Second, from this set of proteins we generated MS queries that closely resembled peptide sequences typically obtained by the interpretation of tandem mass spectra.

We employed random selection of protein sequences to create a dataset that is statistically homogeneous within the entire proteome of a given organism. We note that popular computational methods, such as *bootstrap* and *Monte Carlo* that are very sensitive to representative and unbiased sampling, are also based on random selection of data (35). The software randomly sampled 1,000 proteins from each of three fungal species and 500 proteins from each of the vertebrate species. The sampling strategy was validated on a dataset of *S. cerevisiae* proteins (datasets from other species were built similarly). To this end, we first computed the distribution of the length of proteins within the *S. cerevisiae* dataset, compared it to the distribution of the length of proteins in the whole budding yeast proteome, and found that these distributions overlapped within the margin ±3% (Fig. 1S *A* in the supplemental materials). Next, we performed BLAST2p searches with full-length sequences of proteins from the *S. cerevisiae* dataset against the complete proteome of *C. albicans*. In each search, the top hit was fetched and the percentage of identity of its sequence to the sequence of the queried *S. cerevisiae* protein was calculated. The percentage of budding yeast proteins that share a given percentage of sequence identify with *C. abicans* homologues was plotted. In a separate experiment, all proteins from the proteome of *S. cerevisiae* were searched against the complete proteome of *C. albicans*. The results of the two experiments suggested that the two distributions overlapped within the margin of ±5% (Fig. 1S *B* in the supplemental materials). We therefore concluded that datasets built by random sampling of a large number of proteins reasonably represent physicochemical properties and evolutionary conservation of sequences of proteomes of model organisms.

The number of peptides sequenced *de novo* by tandem mass spectrometry varies greatly between experiments; how-

ever, it rarely exceeds 20 if the amount of analyzed proteins is at the low-picomole level (11). Therefore computations were performed with MS queries comprising from 1 and up to 20 unique peptides, which were randomly fetched from sequences of the protein dataset. Each unique peptide consisted of 10 amino acid residues, which is close to the average length of a tryptic peptide. We further assumed that *de novo* interpretation of tandem mass spectra often does not render fully accurate peptide sequences, but rather yields complete sequence variants in which compatible isobaric combinations of amino acid residues fill sequence gaps. All these sequence variants may be included into the MS BLAST query without affecting the confidence of protein identification, as was explained above. Therefore, to mimic the limited accuracy of *de novo* sequencing, we randomly replaced one or two amino acid residues in each peptide sequence by null-scoring X symbols. This represents a realistic scenario, which assumes that for each fragmented precursor even the most accurate candidate sequence still contains two false amino acid residues (12, 13).

The queries were then used for MS BLAST searches against protein databases of the two remaining species within the same phylogenetic lineage (Fig. 1). To identify homologous proteins in other organisms, we performed BLAST searches with the full-length sequences of the proteins from which MS BLAST queries were generated. In full-length BLAST searches, E-value of 1E-05 was used as a cutoff threshold, so that all hits with higher E-values were disregarded. By re-examining the output of full-length BLAST searches, we estimated that in total more than 90% of hits had E-values lower than 1E-20 and concluded that full-length BLAST hits were statistically significant.

Genes are often multiplied during the evolution. For example, a single gene in *S. cerevisae* might have two or more homologues in *S. pombe* (36), which share substantial identity of their sequences and display similar structural domains. Therefore, if a full-length BLAST search hit more than one protein, confident matching of the corresponding MS BLAST query to any of these proteins was regarded as a positive identification. The percentage of full-length sequence identity between the queried protein and the homologous protein from another organism was collected from full-length BLAST searches performed as described above and was normalized to the length of the queried protein.

MS BLAST hits were regarded as positive only if they met the significance criteria according to the MS BLAST scoring scheme (15, 18). Some typical examples are provided in Table I. In the first example, an MS BLAST query was assembled from peptides from the mouse protein AK002456.1. MS BLAST search identified a *T. rubripes* protein (ID 15391) with six out of eight peptides reported as HSPs. The MS BLAST scoring scheme ignores E-values, *p* values, and bit scores of individual HSPs. Instead, their raw scores are compared with precomputed significance thresholds that are set conditionally depending on the number of aligned HSPs and on the number of unique peptides in the query (see "Experimental Procedures" for details. A full list of precomputed threshold scores is provided in Table 1S in the supplemental materials). First, the score of the top-ranked HSP (in this example, 64) is compared with the threshold score for a single-aligned HSP (59). Because 64 is more than 59, the protein is positively identified and it is not even necessary to consider other re-

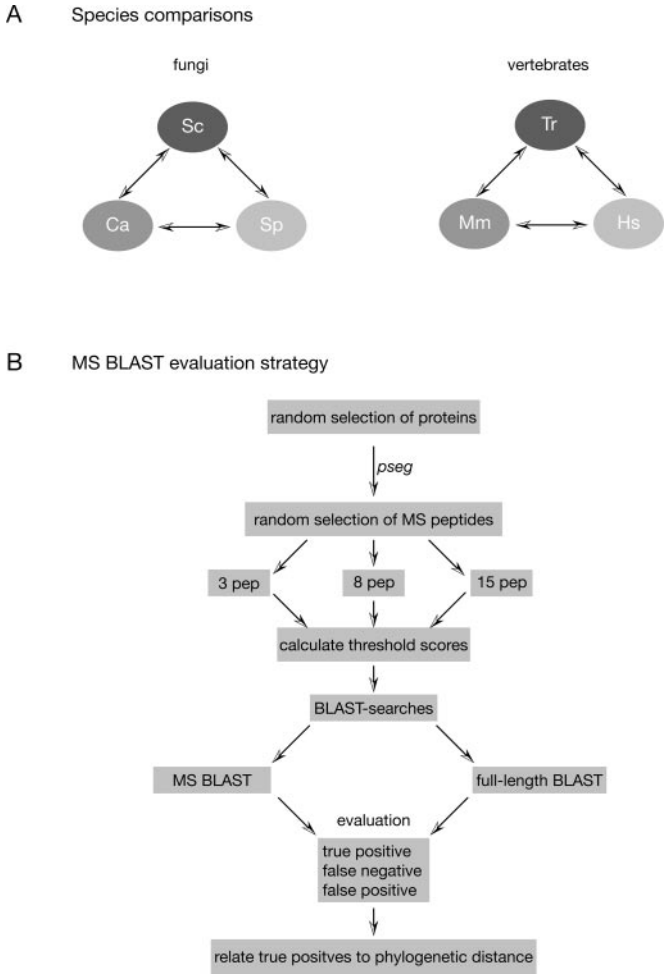

FIG. 1. **Computational strategy for evaluating MS BLAST performance.** *A*, cross-species WU-BLAST2 and MS BLAST searches were carried out for selected species from the fungal and vertebrate lineages. In the case of *S. pombe* (Sp), *T. rubripes* (Tr), *M. musculus* (Mm), and *H. sapiens* (Hs), 3, 8, and 15 peptides were used for MS queries. In the case of *C. albicans* (Ca) and *S. cerevisiae* (Sc), 1, 3, 5, 8, 10, 15, and 20 peptides were used. *B*, a strategy for MS BLAST evaluation. First, the threshold values were calculated based on random alignments of MS queries with sequences from the inverted nonredundant database. Second, MS BLAST and full-length BLAST searches were carried out using randomly sampled proteins from the selected species. Third, the percentage of true-positive, false-negative, and false-positive hits of MS BLAST searches was determined by comparing the hits obtained by MS BLAST to the homologues identified by full-length BLAST searches. Finally, the percentage of true-positive hits was related to the phylogenetic distance between the selected species.

TABLE I

Examples of true-positive, false-negative, and false-positive identifications by MS BLAST

MS BLAST queries were assembled from *M. musculus* proteins and searched against the *T. rubripes* protein database. The precomputed significance threshold scores were: for 1 HSP, 59; for 2 HSPs, 99; 3 HSPs, 131; 4 HSPs, 167 and are reported in bold in the "MS BLAST scoring" column. A full list of threshold scores is provided in the supplemental materials.

Queried protein/ MS BLAST query	MS BLAST hit	HSPs	MS BLAST scoring/result	[a]Full-length BLAST hit (% identity)	Hit
AK002456.1 VNVXVSAEDL- GAFTXXSDFL- XEGDTPRXNK- GVYNXHVXCL- QIRDQXSXGS- VXFGEDIDLP- CEHXVNGXRP- YKXEAGDXMG	ID 15391	Score = 64 Query: 67 CEHXVNGXRP 76 CEH VNG RP Sbjct: 184 CEHHVNGSRP 193 Score = 51 Query: 45 QIRDQXSXGS 54 +IRDQ S GS Sbjt: 96 EIRDQGSCGS 105 Score = 42 Query: 1 VNVXVSAEDL 10 + V SAEDL Sbjct: 127 ISVELSAEDL 136 Score = 41 Query: 12 GAFTXXSDF 20 GAFT DF Sbjct: 249 GAFTVYEDF 257 Score = 39 Query: 34 GVYNXHVXC 42 G Y+ H+ C Sbjct: 168 GLYDSHIGC 176 Score = 30 Query: 78 YKXEAG 83 Y EAG Sbjct: 207 YRCEAG 212	64 > **59** POSITIVE	ID 15391 (61%)	TRUE POSITIVE
AK002456.1 CNKSCXAXYS- AGRNFYXXDI- XSYSVSXSVK- XLGGPKLPGR- EDIDLPXTFD- VNVEXSAEDL- YKHXAGXMMG- LPGXVAFXED	ID 15391	Score = 50 Query: 56 VNVEXSAEDL 65 + VE SAEDL Sbjct: 127 ISVELSAEDL 136 Score = 37 Query: 12 AGRNF 16 AGRNF Sbjct: 41 AGRNF 45 Score = 35 Query: 24 SYSVS 28 SYSVS Sbjct: 227 SYSVS 231 Score = 34 Query: 37 GPKLP 41 GPKLP Sbjct: 63 GPKLP 67	50 < **59** 50 + 37 < **99** 50 + 37 + 35 < **131** 50 + 37 + 35 + 34 < **167** NEGATIVE	ID 15391 (61%)	FALSE NEGATIVE
BAB31737.1 WXTFGLTDTN- XPLSCSLLLV- XTGXLGLNLA- XQLITQAKQT- GPMXKLVXKL- LFVSFLXRAL- SFLNRAXRTD- XQLTLALXSA	ID 13094	Score = 60 Query: 56 LFVSFLXRAL 65 LFVSF+ RA+ Sbjct: 137 LFVSFILRAI 146	60 > **59** POSITIVE	No hit	FALSE POSITIVE

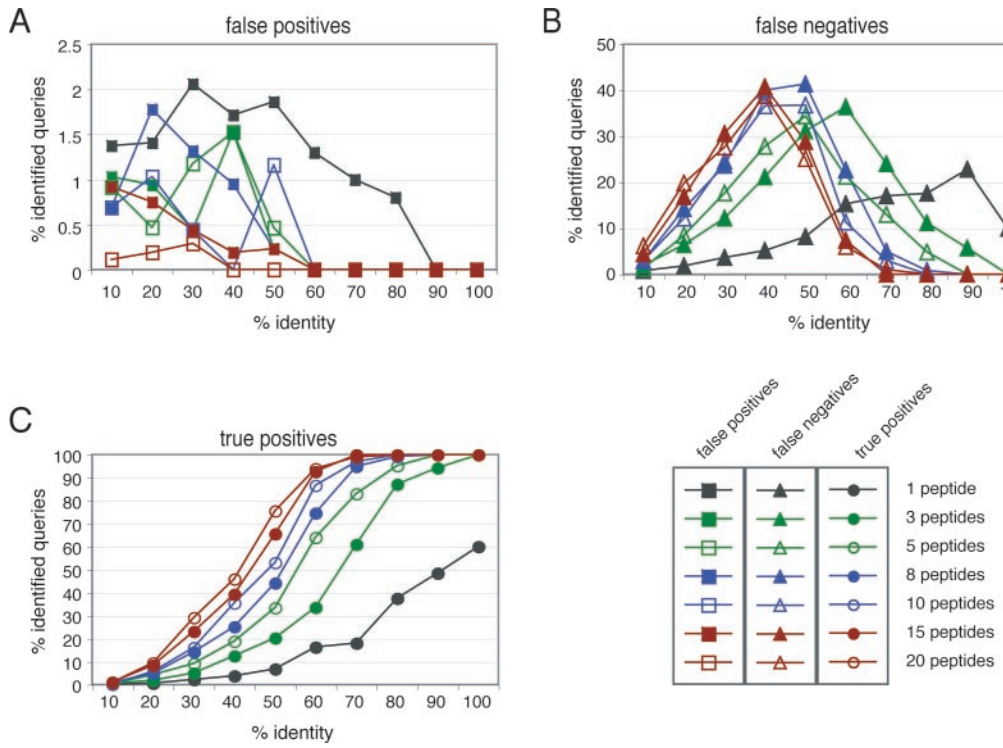

FIG. 2. **Distribution of false-positive, false-negative, and true-positive identifications by MS BLAST.** The percentage of MS BLAST queries identified as false-positive, false-negative, and true positive hits (at the *y*-axes) was related to the percentage identity between the sequence of the protein from which the MS BLAST query was composed and the sequence of its homologue determined by the full-length BLAST search (at the *x*-axes). Calculations were performed for different numbers of peptides in MS BLAST queries. *A*, false-positive hits of MS BLAST searches. *B*, false-negative hits of MS BLAST searches. *C*, true-positive hits of MS BLAST searches. The results of searches with queries composed from *S. cerevisiae* peptides against the *C. albicans* proteome are presented.

ported HSPs. In the second example in Table I, a query composed from another selection of peptides from the same mouse protein hit the same protein from *T. rubripes*. The score of the top HSP (50) was lower than the threshold score for a single-aligned HSP (59). Therefore, the alignment of the two top-scoring HSPs was considered. Their additive score ($50 + 37 = 87$) was also lower than the threshold score for two-aligned HSPs (99). Because the additive score of three HSPs ($50 + 37 + 35 < 131$) and of four HSPs ($50 + 37 + 35 + 34 < 167$) also did not exceed the corresponding thresholds, the identification was considered negative in this case. Following the same scoring scheme, in the third example, a query of eight peptides from the mouse protein BAB31737.1 positively identified the protein ID 13094 by a single-reported HSP ($60 > 59$), albeit no other peptides were aligned. We note that at the MS BLAST web interface (see above) the same scoring procedure is performed by a special script.

We further compared the results of MS BLAST identification with the results of full-length BLAST searches. A BLAST search with the complete sequence of the protein AK002456.1 also identified the protein ID 15391, and 61% of the sequence identity was reported. The MS BLAST identification in the first example was therefore considered as "true positive" because the protein ID 15391 was identified by both MS BLAST and full-length BLAST searches. The second example was considered as "false negative" because the full-length BLAST identified a homologous protein in a database, but MS BLAST failed to do so. In the third case, MS BLAST confidently identified ID 13094 in the database, but this protein was not among the hits of the full-length BLAST search, and this MS BLAST identification was considered as a "false positive."

We investigated the relationship between the rate of true-positive, false-negative, and false-positive identifications by MS BLAST, overall sequence identity between homologous proteins and the number of peptides in a query (Fig. 2). Only a minor fraction of hits were false positives, typically not exceeding 3% of all queried proteins (Fig. 2*A*), well in agreement with the 1% false positives anticipated from the method of calculation of threshold scores. Most false-positive hits were observed in queries with a small number of peptides, and their percentage dropped with increasing the sequence identity between the query and the database sequence. The percentage of false-negative hits was typically in the range between 0 and 40%, and this number decreased with increasing the sequence identity or the number of queried peptides (Fig. 2*B*). Even though the maximal rate of false-negative hits seems quite high, it is not surprising considering that the

percentage of identical residues was calculated based on the entire length of the queried protein. Two proteins might share a single domain or display high enough similarity to be identified by a full-length BLAST search. In case of only local sequence similarity, normalization of the percentage of identical residues to the length of the query lowers the overall identity considerably. In MS BLAST searches, peptides in the query might, for instance, not coincide with the region of similarity between the query and the hit. With increasing the number of unique peptides in the query, the sensitivity of MS BLAST searches almost reaches the sensitivity of full-length BLAST searches for proteins sharing substantial sequence identity with their homologues in a database. While the peak of false negatives could be at as high as 90% of sequence identity for queries comprising only one peptide, it peaked at roughly 40% identity when 20 input peptides were used. The percentage of true-positive hits grew steadily with increasing the percentage of sequence identity or the size of the query (Fig. 2*C*). Using three peptides as an input, 100% of proteins could only be identified when they shared between 90 and 100% sequence identity, and 60% sequence identity was required to identify more than half of the input queries. In case of eight peptides, 50% sequence identity was sufficient to identify over 50% of input queries. It is therefore safe to assume that very few hits will be missed by MS BLAST searches, once the sequence identity to a homologue in a database exceeds 60%. Although this estimate might not look very exciting, we note that it is well beyond the reach of stringent database searching, because on average one out of three amino acid residues in the protein sequence is expected to differ from the sequence of a homologous protein. We noted that the success rate of MS BLAST identification almost reached its maximum when 15 peptides were assembled in the query, and further increase in the number of sequenced peptides (for example, up to 20 peptides) did not enhance its performance substantially.

We next asked whether the percentage of true-positive, false-negative, and false-positive hits in MS BLAST searches would differ depending on the proteome-wide sequence similarity between organisms. To this end, we selected 500 proteins each from the vertebrates *T. rubripes*, *M. musculus*, and *H. sapiens* and repeated the computer simulation experiments as described above. As shown in Fig. 3, *A–D*, the most notable difference is the number of false-negative hits. While the percentage of false negatives never exceeded 20% in the vertebrate lineage (Fig. 3, *C* and *D*), it was twice as high among the fungal species (Fig. 3, *A* and *B*). This agrees with the difference in the overall similarity between the selected proteomes. The number of proteins with less than 40% sequence identity between human and mouse was, for instance, considerably smaller as compared with the fungal species (see Fig. 5*A*) and correlated with the observed rate of false-negative identifications. Furthermore, the similarity of closely related proteins between human, mouse, and *T. rubripes* more likely covers the entire sequence, while in the fungal lineage it is often restricted to a segment of the full sequence.

MS BLAST Searches at the Proteome Scale—We next wanted to know what fraction of the proteomes of the selected species could be identified by MS BLAST, irrespective of the rate of divergence between homologous sequences. We therefore calculated the percentage of true-positive identifications by MS BLAST, depending on the number of unique peptides in the input query. Within the fungal lineage (*S. pombe*, *C. albicans*, and *S. cerevisiae*), MS BLAST could identify less than 30% of queried proteins, even when 15 unique peptides were used as an MS query (Fig. 4, *A* and *B*). The MS BLAST success rate was significantly higher for vertebrates (*T. rubripes*, *M. musculus*, and *H. sapiens*) (Fig. 4, *C* and *D*). Using as few as three peptides per query, MS BLAST could match over 60% of *M. musculus* proteins to human sequences (Fig. 4*D*), and over 80% of queries could be identified with 15 queried peptides. Still, 50% of true positives were found when mouse MS queries were searched against the *T. rubripes* proteome (Fig. 4*C*). Our simulations suggested that within the vertebrate, or rather the mammalian subkingdom, entire proteomes could be covered by MS BLAST. We speculate that the identification of a variety of mammalian proteins might not require any further knowledge of genomes, but can be attained on the basis of already available sequence resources, yet the completeness of the annotation of proteomes is undeniably an important factor. Searching against a database of mouse proteins, the percentage of true-positive human hits dropped below 50% even with 15 unique peptides in the query (Fig. 4*E*), as compared with the over 80% in the reverse direction, apparently because the current database of mouse proteins is less complete than the one from humans. A considerable fraction of human proteins cannot be identified by sequence similarity searches against mouse, because the murine homologues are currently absent in the mouse protein database. Improved annotation of genomic sequence would likely solve this problem in the near future. A majority of missing proteins could still be identified by MS BLAST searches against expressed sequence tag databases. MS BLAST could be applied using the tBLASTn program to search DNA databases. Because sequence matching in this case also relies on protein sequences, the threshold scores obtained for peptide against protein matching will be valid. At the same time, MS BLAST performance within the fungal lineage was less encouraging and it could not be improved substantially. Our simulations suggested that a vast majority of true-positive hits were matched to proteins from the most related species. For example, the success rate was almost unchanged when queries from *C. albicans* proteins were searched against the closely related organism *S. cerevisiae* or against the complete nonredundant database (data not shown). Therefore, enlarging the size of a database by merging sequences from distantly related species does not compensate the lack of proteins from closely related species.

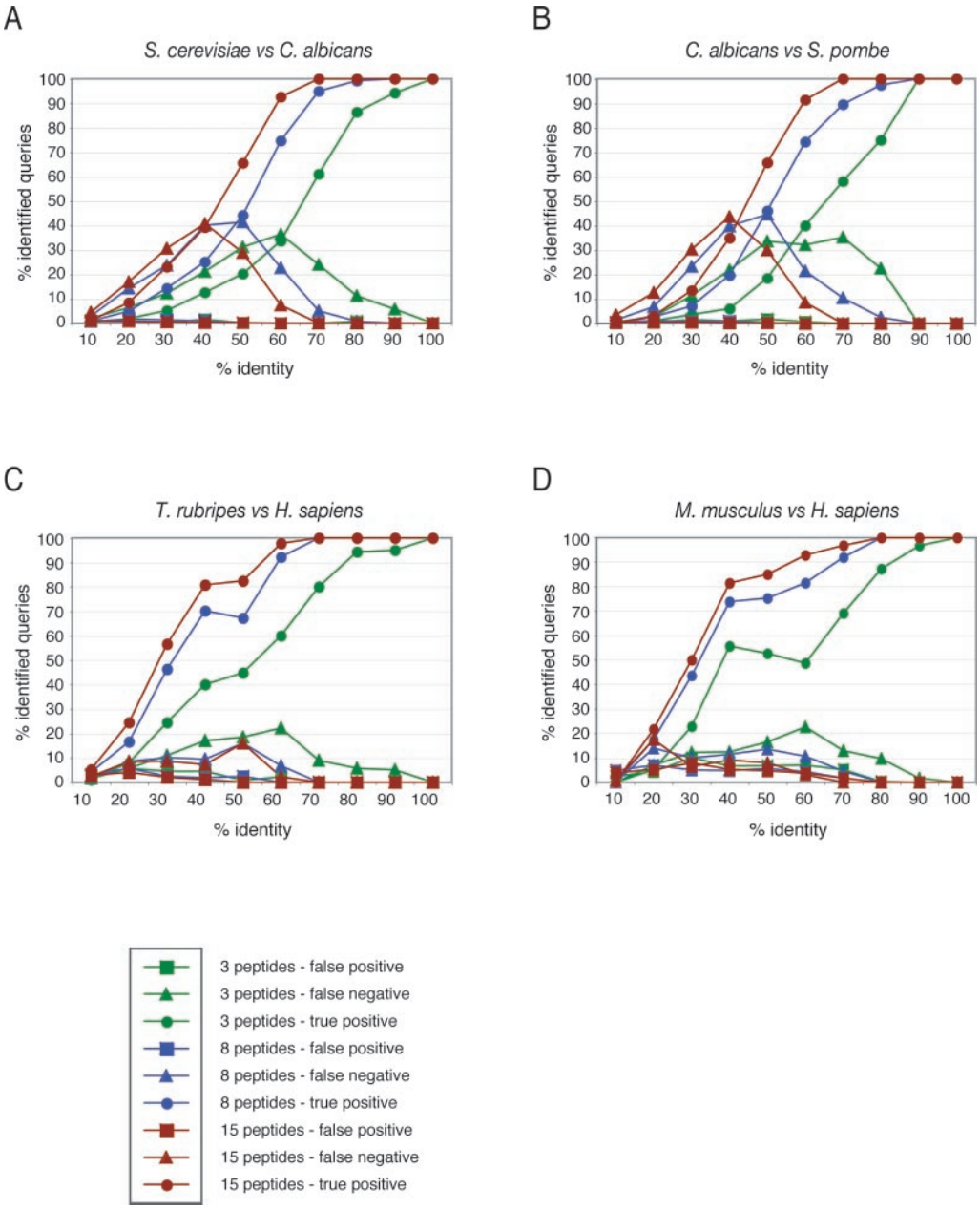

FIG. 3. **Comparative analysis of the percentage of false positives, false negatives, and true positives for different combinations of species.** *A*, *S. cerevisiae* searched against *C. albicans*. *B*, *C. albicans* searched against *S. pombe*. *C*, *T. rubripes* searched against *H. sapiens*. *D*, *M. musculus* searched against *H. sapiens*. The major difference is in the percentage of false negatives in fungal and vertebrate searches: the number of false negatives in the fungal lineage is twice as high as is found between vertebrate species. *Squares*, false positives; *diamonds*, false negatives; *circles*, true positives; *green*, 3 peptides; *blue*, 8 peptides; *red*, 15 peptides.

Prediction of the Success Rate of MS BLAST Identification Based on Phylogenetic Distances—We further estimated which proteomes could be covered by MS BLAST in respect to their distance to the closest related organism having a completely sequenced genome. Phylogenetic analysis is usually based on multiple sequence alignments of sequences of a conserved RNA or protein family, such as mitochondrial small-subunit ribosomal RNA or cytochrome *c* (37), both of which are generally available for a wide range of species. The phylogenetic position of studied organisms relative to a reference organism with a complete genome could help to predict an average success rate of a proteomic study. Another way of estimating the rate of divergence between proteomes is to determine the amount of dissimilar sequences they contain. In other words, the higher the average dissimilarity between proteins of two organisms, the more their proteomes would have diverged. We related the overall success rate of protein identification by MS BLAST to both the percentage of

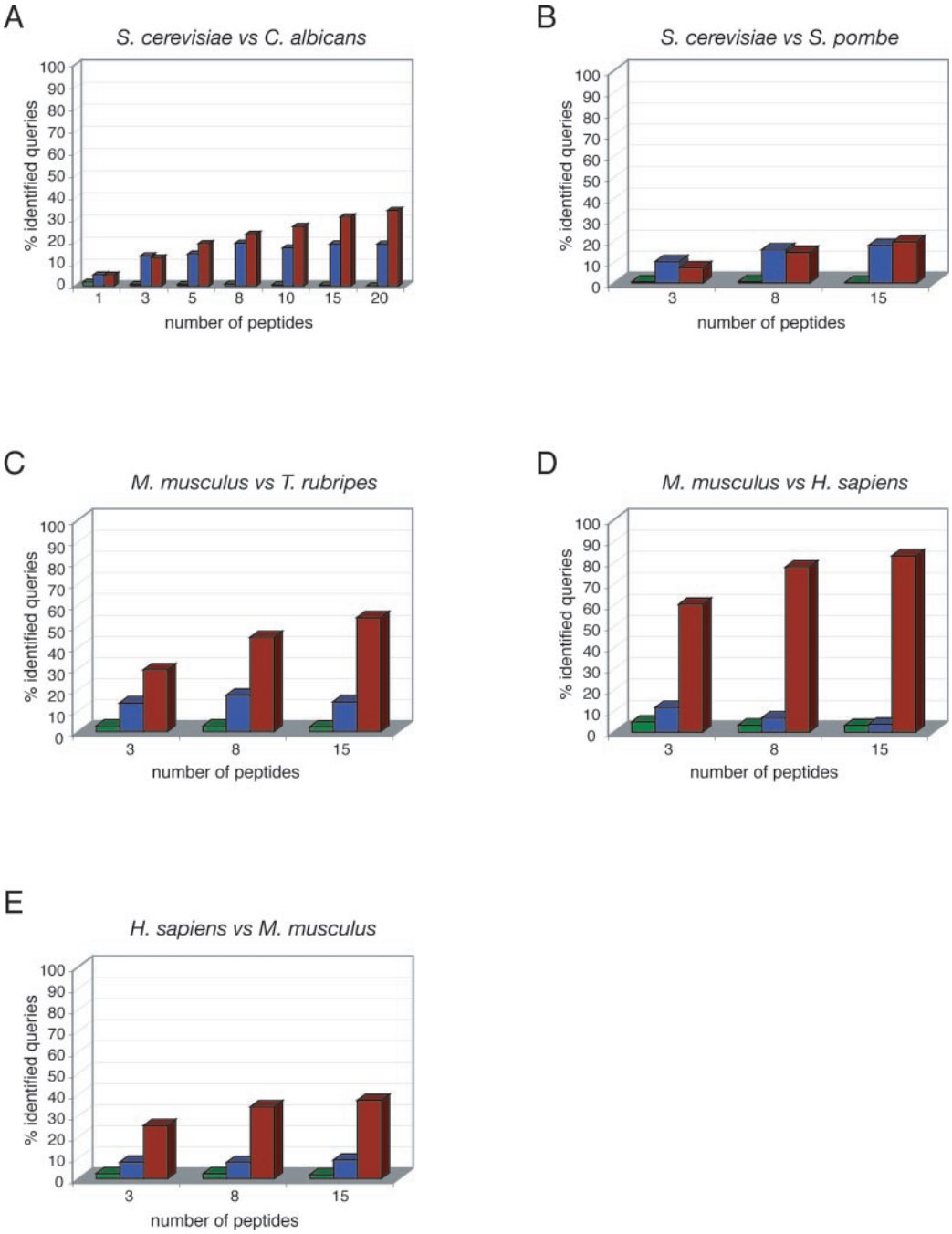

FIG. 4. **The performance of MS BLAST searches in cross-species identification of proteins.** The number of peptides was related to the percentage false-positive (*green*), false-negative (*blue*), and true-positive (red) hits identified by MS BLAST. *A*, *S. cerevisiae* searched against *C. albicans*. *B*, *S. cerevisiae* searched against *S. pombe*. *C*, *M. musculus* searched against *T. rubripes*. *D*, *M. musculus* searched against *H. sapiens*. *E*, *H. sapiens* searched against *M. musculus*.

dissimilar sequences between studied organisms and to the phylogenetic distance based on the alignment of mitochondrial small-subunit ribosomal RNAs. To estimate the divergence of proteomes, we calculated the percentage of dissimilar sequences (*i.e.*, below 20% sequence identity) present in our dataset (Fig. 5*A*). The mouse and human proteomes diverged the least (containing less than 10% of dissimilar sequences), while for the fungal species, nearly half of the proteins in our dataset were considered dissimilar. For example, in the yeasts *S. cerevisiae* and *S. pombe* over 60% of the selected sequences fell into this category. The correlation between the MS BLAST success rate and the divergence of proteomes showed an exponential drop from ~80% of identified proteins, in the case of mouse and human, down to less

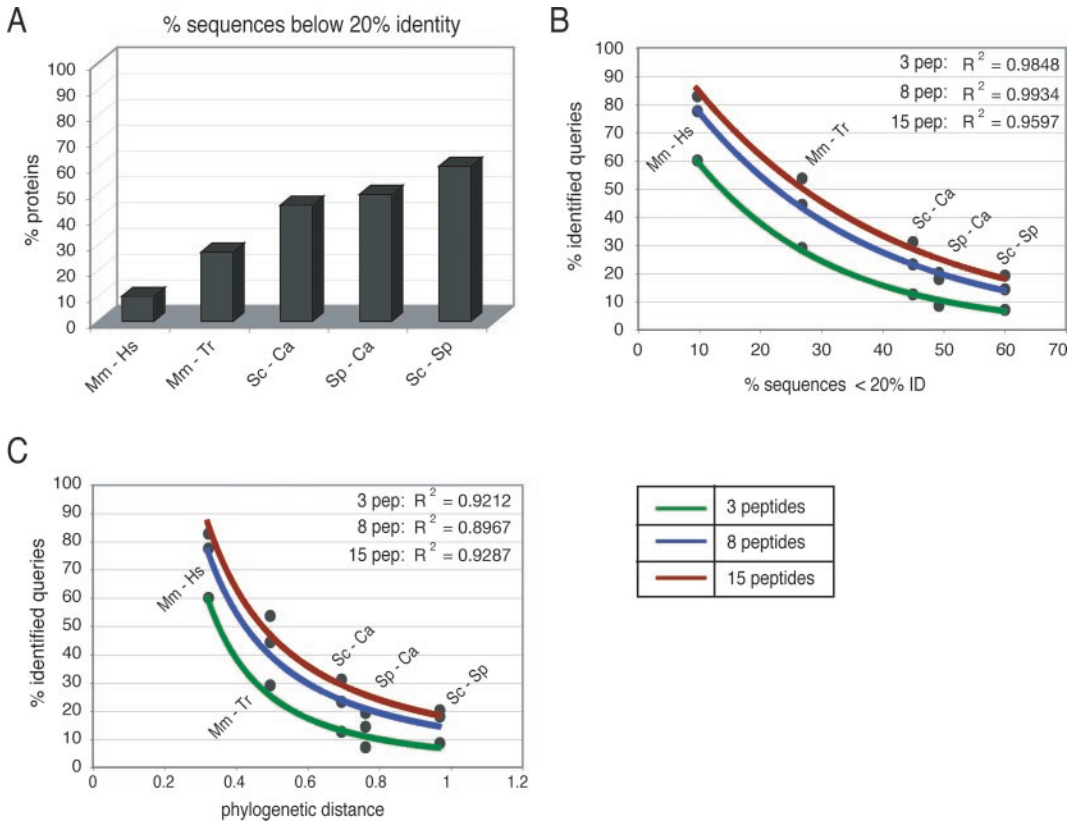

FIG. 5. **Relationship of true-positive identifications by MS BLAST to the phylogenetic distance between species.** The percentage of true-positive identifications by MS BLAST searches depends on the distance to the next completely sequenced genome. *A*, the divergence of proteomes is estimated based on the percentage dissimilar sequences between selected species. *B*, correlation of overall success rate of MS BLAST searches to the percentage of dissimilar sequences in the data set. *C*, correlation of overall success rate of MS BLAST searches to the phylogenetic distance between selected organisms based on the mitochondrial small-subunit ribosomal RNA. The R^2 value of trend line fitting is indicated.

than 30% for the budding and fission yeasts (Fig. 5*B*). Yet for many organisms the paucity of available sequences would not enable calculating the distance to their closest neighbor with a completely sequenced genome using the divergence of their proteomes. In these cases, distances between neighboring organisms can be estimated by the phylogenetic analysis of ribosomal RNA or cytochrome *c*. We performed the phylogenetic analysis of the mitochondrial small-subunit ribosomal RNA family for the fungal and vertebrate lineages. We estimated phylogenetic distances between the model species and related them to the proteome-wide success rate of MS BLAST identification (Fig. 5*C*). Based on phylogenetic analysis of mitochondrial 12S rRNA, the human genome is the closest completed genome to the *M. musculus* genome, with a phylogenetic distance of 0.32. Within this range, more than 60% of proteins could be identified by MS BLAST with as few as three peptide sequences and over 80% with 15 peptide sequences in a query. At the same time, sequence-similarity identification between organisms with a distance around 1.0 became problematic. The selected fungal species *C. albicans*, *S. cerevisiae*, and *S. pombe* all have a distance close to 1.0, and, on average, only 30% of their proteins could be matched even with 15 peptides in a query. Because increasing the number of sequenced peptides over 15 did not improve the success rate, we reasoned that the phylogenetic distance of ~0.5 represents a reasonable limit for reliable coverage of at least 50% of an unknown proteome. However, this estimate should be treated with caution, as the success rate computed over a whole proteome may not apply to identifying members of different protein families, because their conservation strongly varies.

Statistically confident identification of a protein by mass spectrometry does not necessarily imply the direct, unambiguous, and accurate assignment of the biological function. Proteins are typically identified by matching a few MS/MS peptide spectra to the protein sequence in a database, and no credible information on peptides that happened to escape the fragmentation is provided. Often mass spectrometric identification can only point to a gene (or to a family of related genes) within the same organism, rather than identifying a unique protein product (38). This is also true for sequence similarity identifications. A homology-based identification only implies that statistically significant similarity between peptide sequence(s) in the query and sequence(s) in a database has

been registered. However, the functional significance of this observation depends on many indirect factors, such as the number of matched peptides or the functional diversity of proteins that were identified with the same or similar peptide sequences. Although the identification helps to formulate a plausible working hypothesis on what the function of the protein might be, the ultimate proof always rests with a clear biological experiment (39). We note that even if full-length sequences of homologous proteins would become available (for example, from a cloning experiment), the identity of their function still might not be confidently established (36, 40)

CONCLUSION AND PERSPECTIVES

Recent developments in bioinformatics and mass spectrometry effectively argue against the common notion that the availability of the complete genome of an organism is an ultimate prerequisite for successful characterization of its proteome by mass spectrometry. Sequence similarity searches extend the scope of proteomics beyond the boundaries of genomic sequencing, bridging gaps between organisms with rich sequence information. Within the mammalian subkingdom, over 80% of proteins could be positively identified by sequence similarity searches, because orthologous proteins share substantial sequence identity. Considering the phylogenetic relatedness between vertebrates and the availability of the human, mouse, fugu, and zebrafish genomes, it should already be possible to cover most proteomes in this lineage. The availability of the genomes of *Arabidopsis thaliana*, *Zea mays*, *Oryza sativa*, and *Tritium aestivum*, currently being sequenced, will advance proteomics in many economically important plants. However, the success rate of sequence similarity searches will be inevitably smaller for earlier diverged lineages, such as fungi, and further genomic or expressed sequence tag sequencing efforts will be required for mining their proteomes by mass spectrometry.

MS BLAST, as opposed to FASTS and FASTF, maximizes the raw score, rather than minimizing the smallest sum probability of an alignment. Contrary to the calculated E-value of alignments, the raw score used by MS BLAST is not affected by redundant and/or false peptide sequences. MS BLAST enables high-throughput identification of "unknown" proteins by tandem mass spectrometry, because very accurate interpretation of spectra and hand-picking of reliable peptide sequences is no longer absolutely required. However, accurate *de novo* sequencing will remain important for cloning of new genes (41) that do not have close homologues in a database.

How might the power of search algorithms like MS BLAST be extended to more distantly related organisms? It might be possible to either deduce longer sequence stretches by fragmenting multiply charged ions of large protein fragments or even intact proteins in a top-down approach (reviewed in Ref. 42) or to increase the number of sequenced peptides using advanced LC MS/MS combinations (43). Recent progress in Fourier transform mass spectroscopy technology (reviewed in Ref. 4) has had strong impact on the performance of both top-down and bottom-up protein characterization approaches in proteomics, and we might also anticipate that proteomes of species phylogenetically distant from organisms with completely sequenced genomes could be characterized with high sensitivity and throughput.

Acknowledgments—We thank David Drechsel, Wolfgang Zachariae, Judith Nicholls (Max Planck Institute of Molecular Cell Biology and Genetics, Dresden, Germany), and Toby Gibson (European Molecular Biology Laboratory, Heidelberg, Germany) for critical reading of the manuscript and useful discussions.

* The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

[S] The on-line version of this article (available at http://www.mcponline.org) contains supplemental materials.

§ Current address: Scionics Computer Innovation GmbH, Pfotenhauerstrasse 110, 01307 Dresden, Germany.

¶ To whom correspondence should be addressed: Max Planck Institute of Molecular Cell Biology and Genetics, Pfotenhauerstrasse 108, 01307 Dresden, Germany. E-mail: habermann@mpi-cbg.de or shevchenko@mpi-cbg.de.

REFERENCES

1. Griffin, T. J., and Aebersold, R. (2001) Advances in proteome analysis by mass spectrometry. *J. Biol. Chem.* **276,** 45497–45500
2. Mann, M., Hendrickson, R. C., and Pandey, A. (2001) Analysis of proteins and proteomes by mass spectrometry. *Annu. Rev. Biochem.* **70,** 437–473
3. Yates, J. R., 3rd (2000) Mass spectrometry. From genomics to proteomics. *Trends Genet.* **16,** 5–8
4. Aebersold, R., and Mann, M. (2003) Mass spectrometry-based proteomics. *Nature* **422,** 198–207
5. Liska, A. J., and Shevchenko, A. (2003) Combining mass spectrometry with database interrogation strategies in proteomics. *Trends Anal. Chem.* **22,** 291–298
6. Fenyo, D. (2000) Identifying the proteome: Software tools. *Curr. Opin. Biotechnol.* **11,** 391–395
7. Clauser, K. R., Baker, P., and Burlingame, A. L. (1999) Role of accurate mass measurement (+/- 10 ppm) in protein identification strategies employing MS or MS/MS and database searching. *Anal. Chem.* **71,** 2871–2882
8. Liska, A. J., and Shevchenko, A. (2003) Expanding organismal scope of proteomics: cross-species protein identification by mass spectrometry and its implications. *Proteomics* **3,** 19–28
9. Shevchenko, A., Chernushevic, I., Wilm, M., and Mann, M. (2002) "De novo" sequencing of peptides recovered from in-gel digested proteins by nano-electrospray tandem mass spectrometry. *Mol. Biotechnol.* **20,** 107–118
10. McNagny, K. M., Petterson, I., Rossi, F., Flamme, I., Shevchenko, A., Mann, M., and Graf, T. (1997) Thrombomucin, a novel cell surface protein that defines thrombocytes and multipotent hematopoetic progenitors. *J. Cell Biol.* **138,** 1395–1407
11. Lingner, J., Hughes, T. R., Shevchenko, A., Mann, M., Lundblad, V., and Cech, T. R. (1997) Reverse transcriptase motifs in the catalytic subunits of telomerase. *Science* **276,** 561–567
12. Chen, R. H., Shevchenko, A., Mann, M., and Murray, A. W. (1998) Spindle checkpoint protein Xmad1 recruits Xmad2 to unattached kinetochores. *J. Cell Biol.* **143,** 283–295
13. Aigner, S., Lingner, J., Goodrich, K. J., Grosshans, C. A., Shevchenko, A., Mann, M., and Cech, T. R. (2000) Euplotes telomerase contains an La motif protein produced by apparent translational frameshifting. *EMBO J.* **19,** 6230–6239
14. Taylor, J. A., and Johnson, R. S. (2001) Implementation and uses of auto-

mated de novo peptide sequencing by tandem mass spectrometry. *Anal. Chem.* **73,** 2594–2604
15. Shevchenko, A., Sunyaev, S., Loboda, A., Shevchenko, A., Bork, P., Ens, W., and Standing, K. G. (2001) Charting the proteomes of organisms with unsequenced genomes by MALDI-Quadrupole Time-of-Flight mass spectrometry and BLAST homology searching. *Anal. Chem.* **73,** 1917–1926
16. Huang, L., Jacob, R. J., Pegg, S. C., Baldwin, M. A., Wang, C. C., Burlingame, A. L., and Babbitt P. C. (2001) Functional assignment of the 20S proteasome from *Trypanosoma brucei* using mass spectrometry and new bioinformatics approaches. *J. Biol. Chem.* **276,** 28327–28339
17. Mackey, A. J., Haystead, T. A. J., and Pearson, W. R. (2002) Getting more from less: Algorithms for rapid protein identification with multiple short peptide sequences. *Mol. Cell. Proteomics* **1,** 139–147
18. Shevchenko, A., Sunyaev, S., Liska, A., Bork, P., and Shevchenko, A. (2002) Nanoelectrospray tandem mass spectrometry and sequence similarity searching for identification of proteins from organisms with unknown genomes. *Meth. Mol. Biol.* **211,** 221–234
19. Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. (1997) Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res.* **25,** 3389–3402
20. Pearson, W. R., and Lipman, D. J. (1988) Improved tools for biological sequence comparison. *Proc. Natl. Acad. Sci. U. S. A.* **85,** 2444–2448
21. Altschul, S. F., and Gish, W. (1996) Local alignment statistics. *Methods Enzymol.* **266,** 460–480
22. Pearson, W. R., Wood, T., Zhang, Z., and Miller, W. (1997) Comparison of DNA sequences with protein sequences. *Genomics* **46,** 24–36
23. Pevzner, P. A., Mulyukov, Z., Dancik, V., and Tang, C. L. (2001) Efficiency of database search for identification of mutated and modified proteins via mass spectrometry. *Genome Res.* **11,** 290–299
24. Gish, W. (1996) WU-BLAST2.0. blast.wustl.edu
25. Hippler, M., Stauber, E. J., Shevchenko, A., Suemmchen, P., Maroto, F., and Scigelova, M. (2003) De novo sequencing identifies a Fe-deficiency induced protein. *Proc. 51th ASMS Conf. Mass Spectrom. and Allied Topics, Montreal, Canada*, Abstract WPO-254
26. Nimkar, S., and Loo, J. (2002) Application of a new algorithm of automated database searching of MS sequence data to identify proteins. *Proc. 50th ASMS Conf. Mass Spectrom. and Allied Topics, Orlando FL*, Abstract TPL 334
27. Suckau, D., Resemann, A., Schuerenberg, M., Hufnagel, P., Franzen, J., and Holle, A. (2003) A novel MALDI LIFT-TOF/TOF mass spectrometer for proteomics. *Anal. Bioanal. Chem.* **376,** 952–965
28. Schweiger-Hufnagel, U., Lubeck, M., Suckau, D., Muccitelli, H., and Baessmann, C. (2003) Integrating a new peptide de-novo sequencing tool for sophisticated data analysis. *Proc. 51th ASMS Conf. Mass Spectrom. and Allied Topics, Montreal, Canada*, Abstract TPA-001
29. Liska, A. J., Popov, A. V., Sunyaev, S., Coughlin, P., Habermann, B., Shevchenko, A. *et al.* (2004) Homology-based functional proteomics by mass spectrometry: Application to the *Xenopus* microtubule-associated proteome. *Proteomics*, in press
30. Cooper, B., Eckert, D., Andon, N. L., Yates, J. R., and Haynes, P. A. (2003) Investigative proteomics: Identification of an unknown plant virus from infected plants using mass spectrometry. *J. Am. Soc. Mass Spectrom.* **14,** 736–741
31. Wootton, J. C., and Federhen, S. (1996) Analysis of compositionally biased regions in sequence databases. *Methods Enzymol.* **266,** 554–571
32. Thompson, J. D., Gibson, T. J., Plewniak, F., Jeanmougin, F., and Higgins, D. G. (1997) The CLUSTAL X windows interface: Flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res.* **25,** 4876–4882
33. Felsenstein, J. (1993) PHYLIP (Phylogeny Inference Package) v 3.5c. evolution.genetics.washington.edu/phylip
34. Lester, P. J., and Hubbard, S. J. (2002) Comparative bioinformatic analysis of complete proteomes and protein parameters for cross-species identification in proteomics. *Proteomics* **2,** 1392–1405
35. Feller, W. (1966) *An Introduction to Probability Theory and its Applications*. John Wiley & Sons, Inc., New York
36. Aravind, L., Watanabe, H., Lipman, D. J., and Koonin, E. V. (2000) Lineage-specific loss and divergence of functionally linked genes in eukaryotes. *Proc. Natl. Acad. Sci. U. S. A.* **97,** 11319–11324
37. Doolittle, W. F. (1999) Phylogenetic classification and the universal tree. *Science* **284,** 2124–2129
38. Rappsilber, J., and Mann, M. (2002) What does it mean to identify a protein in proteomics? *Trends Biochem. Sci.* **27,** 74–78
39. Roguev, A., Shevchenko, A., Schaft, D., Thomas, H., Stewart, A. F., and Shevchenko, A. (2004) A comparative analysis of an orthologous proteomic environments in the yeasts *S. cerevisiae* and *S. pombe*. *Mol. Cell. Proteomics* **3,** 125–132
40. Rost, B. (2002) Enzyme function less conserved than anticipated. *J. Mol. Biol.* **318,** 595–608
41. Uttenweiler-Joseph, S., Neubauer, G., Christoforidis, S., Zerial, M., and Wilm, M. (2001) Automated de novo sequencing of proteins using the differential scanning technique. *Proteomics* **1,** 668–682
42. Standing, K. G. (2003) Peptide and protein de novo sequencing by mass spectrometry. *Curr. Opin. Struct. Biol.* **13,** 595–601
43. Bruce, J. E., Anderson, G. A., Wen, J., Harkewicz, R., and Smith, R. D. (1999) High-mass-measurement accuracy and 100% sequence coverage of enzymatically digested bovine serum albumin from an ESI-FTICR mass spectrum. *Anal. Chem.* **71,** 2595–2599