ANALYTICAL BIOCHEMISTRY 222, 202–209 (1994)

Amino Acid Analysis and Protein Database Compositional Search as a Rapid and Inexpensive Method to Identify Proteins

Uwe Hobohm, Tony Houthaeve, and Chris Sander
European Molecular Biology Laboratory, Meyerhofstrasse 1, D-69012 Heidelberg, Germany

Received March 28, 1994

The identification of protein samples in minute quantities of protein samples, e.g., from two-dimensional polyacrylamide gel electrophoresis analysis, is an everyday problem in biology laboratories. Here we show that computer-assisted amino acid analysis can fulfill this task. Amino acid analysis data can be used to compare the amino acid composition of an unknown protein with protein compositions in a database (compositional search). Routine amino acid analysis data can, despite a certain margin of error, be used to identify a protein. Compared to protein sequencing, amino analysis is much cheaper, faster, and allows higher sample throughput. Thus, the method may replace protein sequencing as a first attempt in identification, provided a homolog can be found in the database. © 1994 Academic Press, Inc.

Isolation of a protein in small quantities is at present much easier than its identification. A protein may be isolated by 1D or 2D polyacrylamide gel electrophoresis (PAGE)[1] and blotting, by HPLC or capillary electrophoresis, or by its specific association with other molecules, to name the most commonly used methods. However, isolation and purification is one task; identification of biological function is another. A standard approach is to determine a partial sequence on an automated protein sequencer and, with the sequence at hand, perform a database search. On average, the probability to find a homolog for a newly sequenced protein currently is above 50%, with increasing efficiency as a result of the growth of databases (1). However, as more than half of all natural proteins have N-terminii that are chemically modified (2), direct Edman sequence analysis is often not possible.

Automatic amino acid analysis systems perform amino acid hydrolysis, derivatization, and subsequent separation by HPLC almost without manual intervention. The sensitivity is comparable with protein sequencing: analyses with 1–5 μg protein have been done routinely, and analyses down to 0.2 μg, e.g., from 2D PAGE spots, have been done successfully. Here we show that such small quantities of material are in many cases sufficient to identify a protein by amino acid analysis followed by database composition search (PROPSEARCH).

Previous attempts at using amino acid analysis to identify proteins include those of Sibbald *et al.*, Eckerskorn *et al.*, and Shaw (4–6). Sibbald *et al.* and Eckerskorn *et al.* used the amino acid composition together with an error margin, which had to be guessed, as binary filters to scan a composition database (4,5). This approach has the disadvantage that no rank can be calculated. Sometimes many hits are reported without ranking; sometimes no hit can be found. In a more advanced approach, Shaw used a method similar to PROPSEARCH. Unfortunately, amino acid-specific weights were determined from only five proteins, the method was tested on only five known proteins, and protein quantities tested were up to 25 times greater than in this report (6).

MATERIALS AND METHODS

Sample Preparation

All chemicals used in the experiments were of the highest purity possible. Amino acid standard solution was purchased from Sigma. Problott and Prospin were from ABI.

The purification procedure was designed to minimize

[1] Abbreviation used: PAGE, polyacrylamide gel electrophoresis.

0003-2697/94 $6.00

potential interferences in the hydrolysis, derivatization, and separation steps during amino acid analysis. Samples were obtained in different ways. If samples were separated on 1D or 2D PAGE, spots were electroblotted onto a PVDF membrane in transfer buffer [10 mM 3-(cyclohexylamino)-1-propanesulfonic acid, 10% methanol, pH 11] (7,8). The PVDF membrane was washed in water 3 × 1 min, stained for 30 s with 0.1% Coomassie blue R-250/50% methanol/10% acetic acid, and destained for 5 × 1 min in 50% methanol/10% acetic acid. The membrane was then rinsed in water three times and air-dried. Protein bands were excised and further analyzed. If samples were obtained from gel elutions or ion exchange chromatography, then unwanted buffers (such as phosphate buffer or buffer concentrations of more than 0.1 M), trace metal contaminants, and detergents were removed using a Prospin centrifugation unit and the protein was thoroughly washed with 50% methanol (9).

Amino Acid Analysis

The amino acid composition and concentration of the proteins were determined by a single analysis on an Applied Biosystems 420 A/H amino acid analyzer, equipped with an automated hydrolysis unit. The chemical reactions were carried out according to the manufacturer's recommendations, except for the hydrolysis of PVDF blots where the membranes were not covered with a Teflon screen but were transferred to the frit in 30 μl methanol. The hydrolysis conditions were 156°C for 75 min. The hydrolysis peptide standard (CPDFGHIAMELSVRTWKY) was synthesized on an ABI 431A peptide synthesizer using Fmoc chemistry. Calibrations using the standard peptide were done according to the recommendations of the manufacturer and following Bello *et al.* (10).

A Database of Amino Acid Composition

In January 1994 the SwissProt database (11) was updated with about 33,000 protein sequence entries. It was used to precalculate a database of amino acid compositions as follows. SwissProt entries with less than 30 residues were not considered. For each protein in the database, the percentage of 16 amino acids (D + N, E + Q, S, G, H, R, T, A, P, Y, V, M, I, L, F, K) was calculated from the sequence. Asparagine and glutamine are converted to their corresponding acids during hydrolysis and cannot, therefore, be quantified. Tryptophan and cysteine were not considered because they cannot be measured reliably. These 16 numbers, expressed in percentage of the sequence length, were used as amino acid composition. Values were normalized such that 16 amino acids instead of 20 added up to 100%. In a second normalization step, amino acid content values were divided by the standard deviation of the respective database average.

In addition, the $\log_{10}$ of the molecular weight and the calculated isoelectric point were stored as well. We used the logarithm of the molecular weight to have all the values in about the same numerical range. The isoelectric point was calculated as described for the GCG program package (12). Finally, a pointer to the respective SwissProt entry was stored, which was used after a composition search to collect information from the SwissProt database.

The composition database presently has a size of about 2.5 MB. A typical composition search with PROPSEARCH takes a few seconds of CPU time on a common workstation.

Determination of PROPSEARCH Weights for Particular Amino Acid Residues, Molecular Weight, and Isoelectric Point

Amino acids vary in their stability toward acid hydrolysis, oxidation, and derivatization. Systematic amino acid-specific errors in amino acid analysis were calculated by comparing known and measured amino acid composition of 18 proteins, representing proteins of different size (rabbit actin, human serum albumin, aldolase, β-lactoglobulin, bovine serum albumin, che-Y, DNA-K, FTSY, light harvesting complex (LHC), myoglobin, *Aeromonas hydrophila* pro-aerolysin, *Escherichia coli* signal recognition particle 54 kDa (SRP5), mouse SRP5, trypsinogen, tubulin, human Max protein, troponin C, lysozyme). Some proteins were prepared multiple times using different amounts, resulting in 35 amino acid analyses. For the calculation of the amino acid-specific error, only samples with relatively low overall experimental errors of less than 25% were considered. For each analysis, the overall experimental error for amino acids $E_{\text{exp-aa}}$ was calculated as

$$E_{\text{exp-aa}} = \sum_{i=1 \text{ to } 16} |C_i - M_i|,$$

where C is the calculated content of amino acid i and M is the measured content of amino acid i.

The amino acid-specific error E_{aa} was calculated as

$$E_{aa} = C_j - M_j,$$

where C is the calculated content of amino acid j in percentage and M is the measured content of amino acid j in percentage.

Amino acid-specific errors were averaged over 35 samples. Weights were calculated as inverse standard deviation (SD) of mean E_{aa}, resulting in weights between 0.4 and 2.5 (see Table 2). PROPSEARCH weights for iso-

electric point and molecular weight were arbitrarily set to 10 in routine searches, since the experimental determination of isoelectric point and molecular weight is in general more precise than the determination of amino acid content.

Searching for an Unknown Protein in the Database of Amino Acid Composition

Amino acid analysis values for 16 or fewer residue types were calculated and normalized as described above. The approximate molecular weight was known from gel electrophoresis or mass spectrometry and used as additional information.

The distance between the query protein and a protein in the database was calculated as the euclidian distance D_{euclid}

$$D_{\text{euclid}} = (\sum W_i(Q_i - D_i)^2)^{1/2},$$

where W_i is the PROPSEARCH weight, Q_i is the query protein value, measured and normalized, D_i is the database protein value, calculated from the sequence and normalized, i = 1 to 16 is the content of amino acid type i in percentage (entire protein = 100%), i = 17 is the logarithm$_{10}$ of protein molecular weight in Dalton, and i = 18 is the isoelectric point of protein.

If one of these 18 values was not determined, its respective PROPSEARCH weight was set to 0.0. A database search was performed by calculating the distance between the query protein and each protein in the database. Distances were sorted and the smallest 50 distances reported. The protein at the top of the list (rank 1) has the highest probability of being identical to the search protein.

Preparation of a Set of Simulated Amino Acid Analysis Data

One thousand sequences were selected randomly from the SwissProt database. Sequences of length less than 50 residues and more than 1000 residues were excluded, resulting in a collection of 972 sequences. Amino acid content and molecular weight were calculated from the sequences, and a random number generator was applied to introduce deviations from the original values, resulting in simulated amino acid analysis data. The distribution of errors for amino acid content was similar to the distribution of errors in real data (compare Figs. 3B and 4B). The average error for amino acid content was about 14% both in real and simulated data, corresponding to an average error of about 0.9% per amino acid. To simulate an error for the molecular weight, a random number between 0 and 20% of the real molecular weight was generated and randomly added or subtracted from the real molecular weight. The distribution of errors for simulated molecular weight was similar to the distribution of errors in real data.

TABLE 1

Protein Identification by Sequencing or Amino Acid Analysis: Comparison of Cost and Time Effort (3)

	Set-up charge	Per cycle charge	Duration[a]
Protein sequencing	$117	$17	45 min per cycle
Amino acid analysis	$35	—	2 h hydrolysis for all samples plus 45 min per sample

[a] Using automatic analyzers as described under Materials and Methods.

RESULTS AND DISCUSSION

The idea of composition search is simple: identify a protein by comparing its amino acid composition with the composition of all proteins in a protein database. Provided that the amino acid composition has been determined precisely, and the protein or a near homolog has previously been stored in the database, its unique identification is in principle possible. Compared to protein sequencing, which is used extensively to identify proteins, amino acid analysis has a number of advantages: it is cheaper and faster and has a higher sample throughput (see Table 1). So, why is amino acid analysis not a commonly used method to identify proteins ? It has been argued that "amino acid analysis is not as simple and routine as the wide use of the technique may suggest" (13). Different amino acids show different reaction constants during hydrolysis; trace quantities of free amino acids on glassware, in buffers, and even in the acid may introduce background and metals or buffer salts may effect the analysis (14). Strydom *et al.* tried to assess the reliability of amino acid analyses in a collaborative trial among 53 US labs (13). Analyses were performed manually and automatically, with ninhydrin and phenylisothiocyanate (PITC) chemistry. Amino acid-specific errors were not entirely consistent among different methods, but as general trends, the results of Strydom et. al. can be summarized as follows (13):

— Serine and threonine are partly destroyed by hydrolysis (real values higher than measured).

— Tryptophan and cysteine generally show gross errors in many laboratories.

— Glycine values occasionally suffer from contamination, but are reliable if measured by the "best labs."

— Histidine and methionine tend to be comparably problematic, but values from "good labs" again are reliable.

TABLE 2

	Mean difference between measured and calculated amino acid content	SD	PROPSEARCH weight
Asx	1.8	2.44	0.41
Glx	0.0	1.49	0.67
Ser	−0.1	1.39	0.72
Gly	−0.5	1.50	0.67
His	−0.1	0.43	2.33
Arg	0.3	0.52	1.94
Thr	0.3	0.79	1.27
Ala	0.1	0.85	1.17
Pro	0.3	0.58	1.73
Tyr	0.0	0.63	1.58
Val	−1.1	1.30	0.77
Met	0.4	0.66	1.51
Ile	−0.3	0.62	1.61
Leu	0.2	0.76	1.32
Phe	0.3	0.40	2.53
Lys	0.1	0.92	1.09

However, some error remained even in "good" analyses, i.e., amino acid content could be determined only with an error margin of a few percentage points. We will show that a certain error margin is tolerable to identify the protein using PROPSEARCH, and that routine amino acid analysis data lie within this error margin. Here we focus on automatic amino acid analysis using PITC chemistry and automatic vapor-phase hydrolysis. This method is routinely used at the EMBL as a service to in-house and outside groups.

In our experience, a considerable part of imprecise data can be attributed to the chemistry/hardware variability of the instruments. For instance, chemicals and separation columns deteriorate, resulting in a difference for two measurements of test peptide separated by 1 month of up to 10%. Therefore, the routine use of a peptide standard, which corrects for those recovery errors, is essential.

First, we analyzed data on 18 known proteins with respect to amino acid-specific errors. Some proteins were prepared more than once at different times, resulting in 35 analyses (see Table 2). The amino acid-specific error was averaged over 35 analyses. We were not able to confirm the trends reported by Strydom *et al.* (13). In particular, serine and threonine did not give systematically low values. All amino acid-specific errors, including those for glycine, histidine, and methionine, were within one standard deviation of zero (see Fig. 1), indicating a random rather than a systematically low or high deviation. These differences may be attributed to the systematic use of a calibration peptide in our amino acid analyses, which emphasizes the importance of calibration.

To test our composition search program PROPSEARCH, 45 amino acid analyses of known proteins were used in a composition search, including those analyses which were used to determine PROPSEARCH weights. A typical input file is shown in Fig. 2a, a typical PROPSEARCH output is shown in Fig. 2b. PROPSEARCH calculates a distance between the query protein and each protein in the composition database. In the following, *Dist-1* means the weighted euclidian distance between the amino acid composition of the query protein and the *top scoring protein found*, i.e., the protein in the composition database with the smallest distance. As PROPSEARCH weight for a particular amino acid we used the inverse standard deviation of the amino acid-specific error (see Table 2). The PROPSEARCH weight for the molecular weight was arbitrarily set to 10. The isoelectric point was not considered because it had not been determined in all cases.

Surprisingly, the correct protein family was identified in 39 of 45 cases (see Fig. 1). In 23 cases the correct protein sequence was identified. From a biologists point of view interested in function, the identification of the correct protein family is of great help for the choice of further experiments. Therefore, we regarded both the identification of the correct protein or a homolog as protein identification. In 6 of 45 cases, the program failed to identify the correct protein family. Two such analyses were on catalase, where it was known that the samples were several weeks old and not expected to give reliable results for the amino acid analysis. Nevertheless, both samples were included to explore the limits of the method. Indeed, in both cases the experimental error was large: 34.4 and 26.6%, respectively. In one other case (human max protein) the experimental error of 34.4 was unusually high as well. In two other cases (human ca-

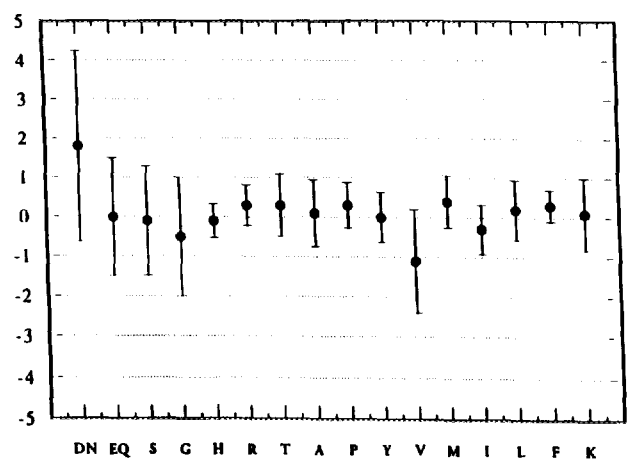

FIG. 1. Amino acid-specific error. Amino acid composition of 18 known proteins has been determined in 35 analyses as described under Materials and Methods. *Y* axis, mean difference between calculated and measured amino acid content (percentage). All amino acid-specific errors were within one standard deviation of zero indicating a random rather than a systematically low or high deviation.

a

#SRP5_ECOLI	
Ala	9.7
Asx	10.4
Glx	9.9
Phe	3.1
Gly	8.5
His	1.1
Ile	4.3
Lys	9.1
Leu	10.4
Met	7.1
Pro	4.3
Arg	5.6
Ser	3.8
Thr	4.7
Val	7.7
Tyr	0.4
MW	49700
pI	4.3

b

Rank	ID	DIST	LEN2	pI	DE
1	SRP5_ECOLI	**1.04**	453	10.30	SIGNAL RECOGNITION PARTICLE PROTEIN (FIFTY-FOUR
2	SRP5_SCHPO	2.07	522	9.82	SIGNAL RECOGNITION PARTICLE 54 KD PROTEIN HOM
3	FLIG_CAUCR	2.32	340	4.69	FLAGELLAR SWITCH PROTEIN FLIG.
4	SRP5_CANFA	2.52	504	9.29	SIGNAL RECOGNITION PARTICLE 54 KD PROTEIN (SRP
5	SRP5_MOUSE	2.54	504	9.47	SIGNAL RECOGNITION PARTICLE 54 KD PROTEIN (SRP5
6	ACCD_ECOLI	2.89	304	8.45	ACETYL-COENZYME A CARBOXYLASE CARBOXYL TRANSFER
7	OPPD_SALTY	2.91	335	6.09	OLIGOPEPTIDE PERMEASE MEMBRANE PROTEIN OPPD.
8	CH60_LEGMI	2.93	546	5.12	60 KD CHAPERONIN (PROTEIN CPN60) (58 KD COMMON
9	ST35_FUSOX	2.93	320	6.18	STRESS-INDUCIBLE PROTEIN STI35.
10	TRPD_METTH	2.94	350	4.56	ANTHRANILATE PHOSPHORIBOSYLTRANSFERASE (EC 2.
11	CH60_LEGPN	2.98	547	5.20	60 KD CHAPERONIN (PROTEIN CPN60) (58 KD COMMON
12	CHEB_ECOLI	3.02	349	8.78	PROTEIN-GLUTAMATE METHYLESTERASE (EC 3.1.1.61).
13	CH60_AMOPS	3.04	551	5.27	60 KD CHAPERONIN (PROTEIN CPN60).
14	THIK_PSEFR	3.05	390	6.99	3-KETOACYL-COA THIOLASE (EC 2.3.1.16) (FATTY O
15	YAAS_ECOLI	3.07	428	5.88	HYPOTHETICAL 45.8 KD PROTEIN IN CARB-KEFC INT
16	LEU3_AGRTU	3.16	370	5.12	3-ISOPROPYLMALATE DEHYDROGENASE (EC 1.1.1.85) (
17	HXKP_RAT	3.19	465	5.00	HEXOKINASE D, PANCREATIC ISOZYMES (EC 2.7.1.1)
18	GAG_FIVPE	3.20	450	8.93	GAG POLYPROTEIN (CONTAINS: CORE PROTEINS P15,
19	YN21_CAEEL	3.23	489	9.90	PUTATIVE ATP-DEPENDENT RNA HELICASE T26G10.1 IN
20	ST35_FUSSH	3.23	324	6.90	STRESS-INDUCIBLE PROTEIN STI35.

FIG. 2. (a) Example for a PROPSEARCH input file. Data files in this format may be sent by electronic mail to hobohm@embl-heidelberg.de for PROPSEARCH analysis. Both upper and lower case characters are accepted, as well as arbitrary (non-zero) number of blanks separating fields. Lines beginning with "#" are comments and skipped. Lines may be in any order. The isoelectric point line or the molecular weight line can be omitted. Also data on amino acids known to be problematic in a particular lab should be omitted (see Results and Discussion). (b) Example for PROPSEARCH output, using the amino acid analysis data for the *E.coli* signal recognition particle protein (SRP5__ECOLI) from Table 2 as query. Identifiers are from the SwissProt database (11). Dist is the distance calculated by PROPSEARCH, Len2 is the number of residues in the sequence found, p*I* the calculated isoelectric point. Dist-1 is the distance of the top-ranking protein found and printed in bold. Protein SRP5 and species *E. coli* were identified correctly at rank 1, with homologous proteins from other species ranking 2, 4, and 5.

thepsin D and DNA K protein from *E.coli*) the experimentally determined molecular weights showed a relatively large error of 15 and 21%, respectively, leading to an uncorrect identification. In a sixth case (triose phosphate isomerase) the correct protein was found at position seven in the PROPSEARCH output. In all six cases of misidentification we found both an experimental error higher than average and a relatively high Dist-1, indicating a possibly unreliable identification.

Overall, we find a correlation between experimental error and Dist-1 (Fig. 3): the higher the experimental error, the higher is Dist-1. In case of an unknown protein sample, Dist-1 may be used to give an indication about the quality of the experimental data and the reliability of the protein identified. From the experimental data it may be concluded that as a rule of thumb a PROPSEARCH distance above 2.5 might indicate unreliability, while a PROPSEARCH distance below 1.5 may indicate good reliability, with a possible "twilight" zone in between (see Fig. 3). This reliability threshold was further explored using simulated amino acid hydrolysis data (see below).

To what extent is the performance of PROPSEARCH affected by the use of amino acid-specific weights ? To study this question, weights for 16 amino acids were set equal to 2, and PROPSEARCH was tested on the same set of 45 amino acid analyses. The program still identified 32 protein families correctly (data not shown), compared to 39 correctly identified protein families using amino acid-specific weights. All 13 misidentified proteins had a large overall error in the amino acid analysis of, on average, 23.2%, compared to the average experimental error of 15.5% found for 45 samples. Therefore, PROPSEARCH weights adjusted for expected experimental error appear to improve the search sensitivity if the experimental error is higher than normal, but are not crucial if the amino acid analysis data are of good quality. Analyses were done with 1–5 μg of protein. In a few cases, even 0.2 μg were enough to identify the protein correctly.

A practical question of reliability arises whenever an experimental amino acid analysis and subsequent PROPSEARCH database search has been performed using a sample of an unidentified protein. Is it possible to estimate from Dist-1 the chance of correct identification? In other words, given a particular Dist-1, what is the ratio of true and false positives in a typical case? Obviously, a data set of 45 protein samples is too small to answer this question. Therefore, a data set of 972 simulated amino acid analyses data was prepared. The 972 real sequences were selected from the SwissProt database, and a random number error generator was applied to generate errors for the amino acid content and the molecular weight (see Materials and Methods). The dis-

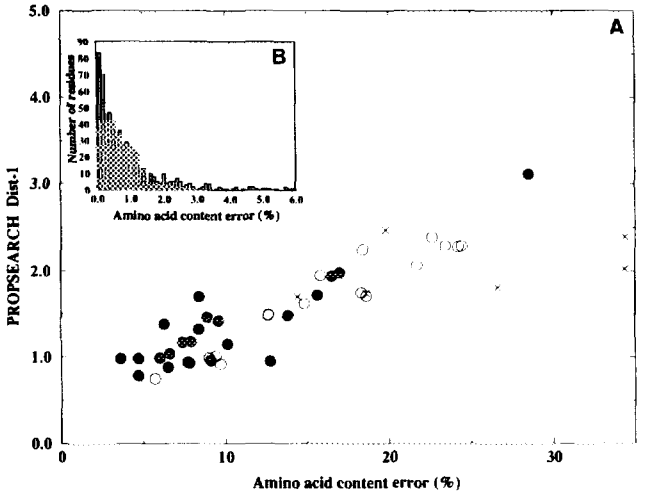

FIG. 3. Correlation between PROPSEARCH distance and amino acid analysis experimental error (A) Forty-five samples of known proteins were characterized by amino acid analysis and a subsequent PROPSEARCH composition database scan. PROPSEARCH distance Dist-1 and experimental error were calculated as described under Materials and Methods. Filled circle, protein *sequence* correctly identified. Open circle, a member of the protein *family* correctly identified. Cross, protein sequence and protein family not correctly identified. A protein was considered identified if it appeared at the top of the list (rank 1). (B) Distribution of amino acid analysis errors.

tribution of amino acid errors was comparable between simulated and real data (compare Figs. 3B and 4B), indicating a realistic distribution of errors in the simulated data set. The distribution of molecular weight errors was comparable between simulated and real data as well (data not shown). Simulated hydrolysis data were analyzed by PROPSEARCH in the same way as real data using PROPSEARCH weights from Table 2 (see Fig. 4A).

Provided that the simulated data sets on average represent real data, one can now calculate the reliability of identification as a function of PROPSEARCH Dist-1 (see Fig. 5). For instance, with a PROPSEARCH Dist-1 of 1.0 and 2.0 the protein family is found with probability 1.0 and 0.68, respectively.

The simulation may not perfectly represent real data sets. For instance, in the real data sets we did not find an incorrect protein family identification below Dist-1 1.7, while in the simulated data the first incorrect protein family appears with Dist-1 1.2. This may be caused by the fact that some real data sets were used to both calculate and test amino acid-specific weights, while the simulated data were independent sets not used for the weight calculation. On the other hand, the reliability thresholds shown in Fig. 5 may represent a worst case estimation, while real experimental data behave more favorably. For instance, in the simulated sets we assumed an average error of about 14 and 10% for amino acid content and molecular weight, respectively. These estimates were derived from real data which were known to be problematic in some cases, but nevertheless considered to explore the limits of the method. Both error margins are smaller if data of known bad quality are rejected. Thus, the reliability of PROPSEARCH must be scrutinized by more real data sets in the future.

If additional biochemical knowledge about the protein under investigation is at hand, e.g., if it is known that the protein is, for instance, of cytoskeletal origin or membrane bound or an enzyme aso, this knowledge may be used to inspect proteins ranking higher than 1 in the PROPSEARCH output. The chance to find the correct protein among the first 20 hits is much higher than to find it on rank 1. For example, with a PROPSEARCH Dist-1 of 2.0 the correct protein family is found with probability 0.68 on rank 1 and with probability 0.92 among rank 1 and 20 (Fig. 5).

Other additional knowledge may be used to adjust the PROPSEARCH weights. If, for instance, the molecular weight has been determined by mass spectrometry to a high degree of accuracy, the PROPSEARCH weight for molecular weight may be increased, while raising the re-

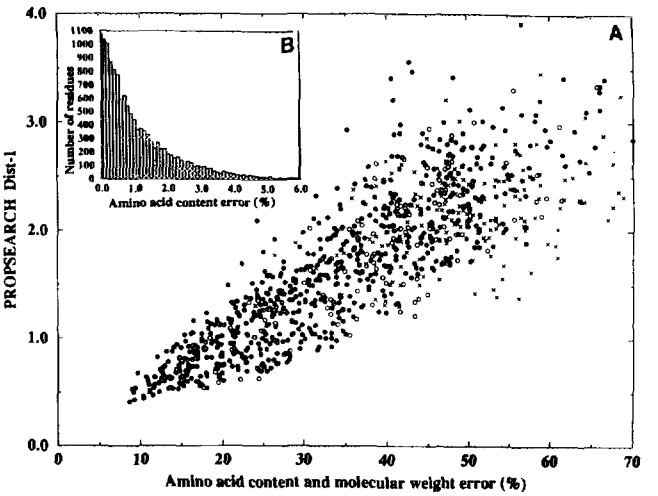

FIG. 4. PROPSEARCH reliability estimation: Analysis of 972 simulated amino acid analysis data sets. (A) A PROPSEARCH composition database search was performed with simulated data sets in the same way as described under Materials and Methods for real data. Summed amino acid errors were allowed up to 25%; errors for the molecular weight were allowed up to 20%. Since the molecular weight error contributes about one-third of the PROPSEARCH distance, the error on the Y axis was calculated as $E_{MW} + E_{aa}{}^{*2}$, where E_{MW} is the molecular weight error and E_{aa} is the error for amino acid content, resulting in a maximal possible error of 70%. Filled circle, protein sequence identified correctly. Open circle, protein family identified correctly. Cross, protein and protein family not identified correctly. A protein was considered identified if it appeared on the top of the list (rank 1). (B) Distribution of amino acid analysis errors in simulated data, for comparison with the distribution of errors for real data (fig. 3b).

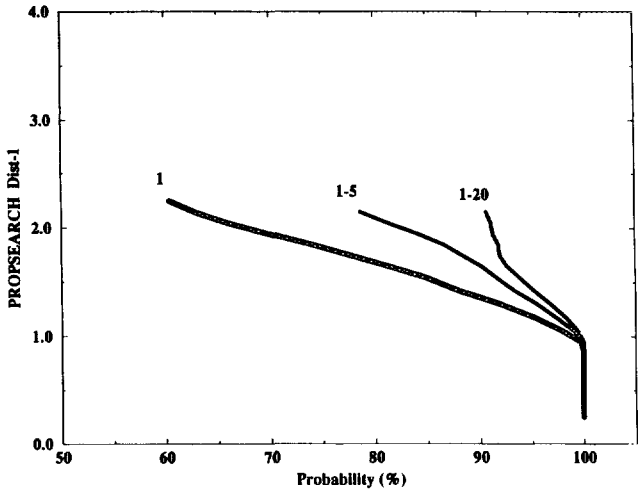

FIG. 5. Probability of protein family identification as assessed by simulated amino acid analysis data. A window of 0.3 PROPSEARCH distance units was shifted along the vertical axis of Fig. 4A in steps of 0.1 PROPSEARCH distance units, and the ratio between the number of correctly identified protein families (filled and open circles) and the number of mis-identified protein families (crosses) inside the window was calculated. Lines from left to right represent the probability of finding the correct protein family on rank 1, 1–5, and 1–20 in relation to PROPSEARCH Dist-1. Note that since data in Fig. 4A are sparse for higher PROPSEARCH distances, lines were plotted only up to the point of dense data. Note also that values were smoothed over four consecutive data points.

liability of identification accordingly. On the other hand, if a post-translational modification is suspected, the PROPSEARCH weight for molecular weight may be decreased, retaining a reliable identification.

When 2D gel electrophoresis has been performed, the experimentally determined isoelectric point can be used as an additional search criterion. To get an indication of the correlation between the experimentally determined isoelectric point and the isoelectric point calculated from the sequence, we compared both values for 37 proteins published by Rasmussen *et al.* (15). To exclude post-translational modifications, which might lead to an imprecise isoelectric point, only those proteins were selected for which calculated and experimental molecular weights were similar. The difference between experimental and calculated isoelectric point was more than one pH unit only in one case, with an average error of 0.3 with a standard deviation of 0.4. As a general rule we therefore assume the calculated isoelectric point to be correct within plus/minus one pH unit. If experimental information about the isoelectric point is available, it is used with a PROPSEARCH weight of 10 in our routine searches.

Sometimes it is known in advance that some amino acid values are particularly inaccurate. For instance, when samples coming from the protein sequencer were further analyzed by amino acid analysis, we found a larger error for glycine, arginine, lysine, and methionine (data not shown). In this case, it was still possible to use PROPSEARCH by omitting those 4 amino acids and using the remaining 12 composition values for the other amino acids. In some cases the correct protein was identified using 12 values only, but not using 16 values including the error-prone ones were used (data not shown). In other words, it is preferable to use fewer, more accurate data points rather than many, less accurate ones.

Of course, multiple amino acid analyses of the same protein will improve certainty further.

CONCLUDING REMARKS

In conclusion, amino acid analysis can in the majority of cases replace protein sequencing as a method to identify an unknown protein, if state-of-the-art experimental and computational methods are used, and if a homologous protein is present in the database.

For the amino acid analysis, it is neccessary that

- the protein sample is free of metal and salt buffer,
- hydrolysis is calibrated with a standard peptide.

The sensitivity of database searches can be improved, if

- particular amino acids are weighted by taking their individual stability toward hydrolysis, derivatization, and separation into account,
- molecular weight is added as an additional data point,
- the isoelectric point is added as another data point.

If an unknown protein is analyzed, then the Dist-1 calculated by PROPSEARCH gives an indication if the identification was reliable, with a small Dist-1 indicating both good quality experimental data and correct identification of the protein or protein family, and with a high Dist-1 indicating either bad quality data or that a homolog is not yet in the database.

With these guidelines in hand, routine use of PROPSEARCH assisted amino acid analysis for protein identification is becoming a practical proposition.

A PROPSEARCH database search can be requested via electronic mail by sending amino acid analysis data to hobohm@embl-heidelberg.de (see Fig. 2a for instruction). To improve statistics, we would like to invite particularly amino acid analyses of *known* proteins.

ACKNOWLEDGMENTS

We thank M. Mann for helpful discussions and for the C version of the subroutine to calculate the isoelectric point, and L. Holm and M. Saraste for critically reading the manuscript. This work was supported by a grant from the German Ministery of Research (BMFT).

REFERENCES

1. Koonin, E., Bork, P., and Sander, C. (1994) *EMBO J.* **13,** 493–503.
2. Aitken, A., Geisow, M. J., Findlay, J. B. C., Holmes, C., and Yarwood, A. (1989) *in* Protein Sequencing: A Practical Guide (Findlay, J. B. C., and Geisow, M. J., Ed.) IRL Press, Oxford.
3. Ivanetich, K. M., Niece, R. L., Rohde, M., Fowler, E., and Hayes, T. K. (1993) *FASEB J.* **7,** 1109–1114.
4. Sibbald, P. R., Sommerfeldt, H., and Argos, P. (1991) *Anal. Biochem.* **198,** 330–333.
5. Eckerskorn, C., Jungblut, P., Mewes, W., Klose, J., and Lottspeich, F. (1988)*Electrophoresis* **9,** 830–838.
6. Shaw, G. (1993) *Proc. Natl. Acad. Sci. USA* **90,** 5138–5142.
7. Vandekerckhove, J., Bauw, G., Puype, M., VanDamme, J., and VanMontague, M. (1985) *Eur. J. Biochem.* **152,** 9–19.
8. Aebersold, R. H., Teplow, D. B., Hood, L. E., and Kent, S. B. H. (1986) *J. Biol. Chem.* **261,** 4229–4238.
9. Sheer, D. (1990) *Anal. Biochem.* **187,** 76–83.
10. Bello, R., Bozzini, M., Chui, A., Noble, R., and Dupont, D. (1989) Proceedings, Third symposium of the Protein Society, Seattle, Washington.
11. Bairoch, A., and Boeckmann, B. (1991) *Nucleic Acids Res.* **19,** 2247–2250.
12. GCG (1991) Program Manual for the GCG Package, Version 7, Genetics Computer Group, 575 Science Drive, Madison, WI 53711.
13. Strydom, D., Tarr, G. E., Pan, Y.-C. E., and Paxton, R. (1992) *in* Techniques in Protein Chemistry, III (Angeletti, R. H., Ed.), Academic Press, New York.
14. ABI (1989) ABI Model 420A User Manual, Version 3.22 .
15. Rasmussen, H. H., VanDamme, J., Puype, M., Gesser, B., Celis, J. E., and Vandekerckhove, J. (1992) *Electrophoresis* **13,** 960–969.