# Getting More from Less

ALGORITHMS FOR RAPID PROTEIN IDENTIFICATION WITH MULTIPLE SHORT PEPTIDE SEQUENCES*[S]

**Aaron J. Mackey‡§, Timothy A. J. Haystead¶‖, and William R. Pearson**‡‡**

**We describe two novel sequence similarity search algorithms, FASTS and FASTF, that use multiple short peptide sequences to identify homologous sequences in protein or DNA databases. FASTS searches with peptide sequences of unknown order, as obtained by mass spectrometry-based sequencing, evaluating all possible arrangements of the peptides. FASTF searches with mixed peptide sequences, as generated by Edman sequencing of unseparated mixtures of peptides. FASTF deconvolutes the mixture, using a greedy heuristic that allows rapid identification of high scoring alignments while reducing the total number of explored alternatives. Both algorithms use the heuristic FASTA comparison strategy to accelerate the search but use alignment probability, rather than similarity score, as the criterion for alignment optimality. Statistical estimates are calculated using an empirical correction to a theoretical probability. These calculated estimates were accurate within a factor of 10 for FASTS and 1000 for FASTF on our test dataset. FASTS requires only 15–20 total residues in three or four peptides to robustly identify homologues sharing 50% or greater protein sequence identity. FASTF requires about 25% more sequence data than FASTS for equivalent sensitivity, but additional sequence data are usually available from mixed Edman experiments. Thus, both algorithms can identify homologues that diverged 100 to 500 million years ago, allowing proteomic identification from organisms whose genomes have not been sequenced.** ***Molecular & Cellular Proteomics 1:139–147, 2002.***

The rapid and accurate identification of proteins from biological isolates is the primary goal of modern proteomics (Fig. 1*A*). Various mass spectrometry techniques, including matrix-assisted laser desorption/ionization-time-of-flight and liquid chromatography-electrospray ionization, can quickly obtain peptide mass mappings that may be matched against theoretical spectra derived from primary sequence databases (1). However, MS[1]-based mass-matching requires a nearly exact match to a database sequence for success, making it impractical for use in organisms whose genomes have not been sequenced to high quality or in the identification of alternatively spliced gene products. More sophisticated algorithms for spectral-based searches are mutation-tolerant (2), but efficiency is maintained only with minimal mutations (2, 3).

Alternatively, tandem mass spectrometry (MS/MS) experiments are capable of generating small amounts of sequence (4); the protein of interest is enzymatically cleaved into multiple peptide fragments that are separated by mass prior to collision-induced dissociation, producing MS/MS spectra of each peptide. Each spectrum can be interpreted to obtain *de novo* partial primary sequence data (5, 6) (Fig. 1*B*), and a sequence-based database search may be performed (7). Previous methods to identify *de novo* peptide sequences perform multiple independent database searches with each peptide sequence and subsequently report the database sequences that occur most often between all of the peptides (8). Although this approach works well for fragments with an exact or near exact match in the database, standard sequence similarity search algorithms have difficulty identifying evolutionary-related sequences sharing less than 90% identity given only a single short peptide (three to six amino acids) because of the limited information content of the search queries (9, 10). Therefore, search results from peptides with lower identity can add more noise than homology signal to the analysis. Here we describe an improved algorithm, FASTS, for database searching with all MS-derived *de novo* peptide sequences simultaneously, under conditions which maximize the information content of the limited query through the use of probability-based scoring.

Partial protein sequences may also be determined by conventional N-terminal Edman sequencing on unseparated mixtures of peptides (11) (Fig. 1*C*). Because no fragment separation step is involved, femtomoles of starting material are sufficient for 10–12 sequencing cycles, where each cycle produces a mixture of all the amino acids present at that cycle from each peptide. Mixed peptide sequencing usually obtains longer sequence reads than that possible by MS-based *de novo* sequencing and is less sensitive to post-translational modifications. However, the exact linear sequence of any individual peptide remains unknown, requiring a deconvolution of the residues at each site to reconstruct the original sequences of each peptide. Previously (11), we briefly described the FASTF algorithm for performing this task; here we provide further details of the FASTF algorithm, as well as improvements to sensitivity

From the Departments of ‡Microbiology and **Biochemistry and Molecular Genetics, University of Virginia, Charlottesville, Virginia 22908 and the ¶Department of Pharmacology, Duke University, Durham, North Carolina 27710

Received, August 7, 2001, and in revised form, November 13, 2001
Published, MCP Papers in Press, December 12, 2001, DOI 10.1074/mcp.M100004-MCP200

[1] The abbreviations used are: MS, mass spectrometry; MS/MS, tandem mass spectrometry.

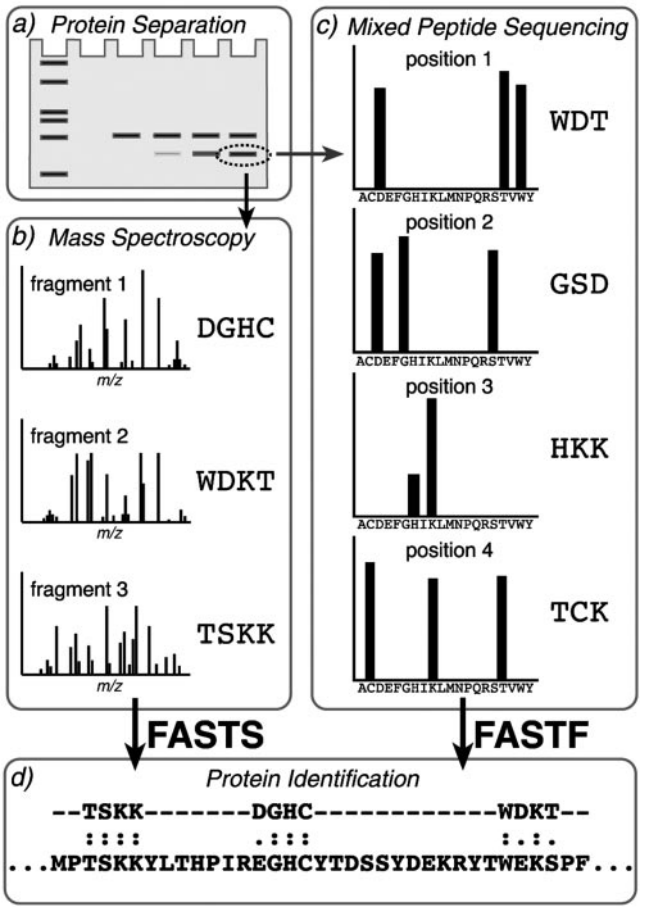


FIG. 1. **Searching with short unordered peptides and peptide mixtures.** FASTS and FASTF find unknown proteins (*a*) using sequence data obtained from MS/MS (*b*) or mixed Edman sequencing (*c*). Mixed peptides sequencing produces multiple residues from each cleavage/analysis cycle. MS/MS peptides are ordered, and peptide mixtures deconvolved and ordered, by mutation-tolerant alignment to related sequences (*d*).

over the previous approach by introducing probability-based scores.

FASTS and FASTF extend the FASTA algorithm (12) and are available in the FASTA software package for sequence database searching. Both algorithms maximize the search sensitivity by (*a*) using scoring matrices with high information content, (*b*) constraining the kinds of alignments generated, and (*c*) using a strict probabilistic criterion for optimal alignments, which significantly improves the sensitivity and specificity of the algorithms over traditional similarity-score maximization approaches. Most importantly, these algorithms calculate accurate statistical estimates, providing the ability to robustly identify homologous proteins from large scale proteomic sequencing efforts.

## EXPERIMENTAL PROCEDURES

*The FASTS and FASTF Algorithms*—FASTS and FASTF use the FASTA heuristic strategy (12) (see Table I) to rapidly search a database for high quality alignments indicative of homology. In the initial stage, regions with high identity to each peptide are identified using a lookup table. These regions are then used as the focal points for generating ungapped subalignments, using a PAM-like (13) scoring matrix (FASTS automatically modifies the scoring matrix to account for Ile/Leu and Lys/Gln isobars). FASTS and FASTF differ from FASTA during this subalignment stage; whereas FASTA looks for the best local ungapped alignment across the region, FASTS and FASTF automatically force an ungapped global alignment of the individual peptide within the library sequence.

In the next stage, FASTA joins subaligned regions, summing their similarity scores to find an optimally scoring, non-overlapping, and appropriately ordered alignment, *i.e.* the regions aligned are linearly ordered within both the query and the library sequences. Because the true order of the query peptides used by FASTS is unknown, FASTS requires only that the aligned peptides do not overlap. Additionally, FASTS joins the subalignments to produce the lowest probability overall alignment, rather than the highest sum of similarity scores (see below under "Alignment Probabilities"). FASTA includes a fourth and final stage to produce a Smith-Waterman alignment with gaps but constrained within a diagonal band centered at the highest scoring initial region. FASTS does not perform this fourth step as gaps are not allowed within the aligned peptides. In summary, FASTS simply extends the FASTA lookup and joining strategy with modifications to align unordered peptides, using a more strict joining criterion.

FASTF uses an identical strategy as FASTS but must also deconvolute the mixture of amino acid residues provided from each cycle of the Edman degradation reaction. This deconvolution requires an additional stage (2b in Table I) to ensure that each residue from each cycle is used only once in the peptide alignments. Because the assignment of residues to peptides is not known in stage 1 and stage 2, FASTF first calculates the best possible similarity scores by selecting the amino acid that produces the highest match score in each position in each high identity region, regardless of whether the selected amino acid had been used previously. For instance, if the residues L, K, and S were given as the amino acids present at position two in the query and were being aligned to library residues L, R, and M at the second stage, then the query residue L would align to both the library L (with a score of +20 using the MDM20 scoring matrix (14)) and the library M (score of −2), and the query residue K would align with the library R (score of 0), whereas the query residue S would remain unused.

During the rescan in stage 2b, the peptide alignments "consume" the best residues available, with the highest scoring region from stage 2a getting to choose first. In the best case, the library residue L should consume the query residue L, because the L:L alignment has the highest score and would be given first choice of the query residues to consume; K should next align to R and then S with M (score of −12). However, if the region containing the library M (the M-region) had a higher overall score than the region containing the library residue L (the L-region), then the M-region would choose its query residues first in stage 2b; the query residue L would be consumed by the M-region before the L-region could use it, forcing the library L to align to a different (and worse-scoring) query residue. Thus, this "greedy" method of deconvolution can generate suboptimal scores that lead to obvious mistakes in the reported alignment (as above where the two pairs L:L and S:M could instead be erroneously aligned as L:M and S:L). The greedy approach may reduce FASTF sensitivity. However, an optimal assignment of residues to peptide alignments is considerably more time consuming because of the combinatorial nature of the problem.

*Alignment Probabilities*—Unlike the similarity scores produced by FASTA, BLAST, and other sequence similarity searching programs that calculate local alignments, FASTS and FASTF calculate align-

TABLE I
*The FASTA, FASTS, and FASTF algorithms*

| | FASTA | FASTS | FASTF |
|---|---|---|---|
| 1. | Identify identical regions shared by query and library sequence with lookup table | Identify identical regions | Identify identical regions with any match in each position |
| 2. | Rescan identical regions using scoring matrix to find best local alignment without gaps | Rescan identical regions requiring global alignment in query | Rescan identical regions selecting best scoring residue at each position |
| 2b. | | | Rescan again, starting with best scoring region, consuming residues from position mixture |
| 3. | Join non-overlapping, ordered regions to produce best scoring alignment | Join non-overlapping regions to produce lowest probability alignment | Join non-overlapping regions to produce lowest probability alignment |
| 4. | Calculate band-limited Smith-Waterman score | | |

ments whose scores are a combination of both global and local similarity scores. Global scores are calculated for individual peptide alignments, but peptides may be left out of the final alignment; a local alignment of the set of globally aligned segments lets contaminating peptides be excluded. These hybrid global/local alignment scores are not extreme value-distributed as are conventional BLAST, FASTA, and Smith-Waterman similarity scores (15, 16) (data not shown). To estimate the statistical significance of an alignment, we first calculate a theoretical probability for it, assuming that it was obtained by an optimal algorithm employing no heuristics. This probability is subsequently scaled to reflect the empirically observed distribution of alignments.

The statistical expectation of a FASTS or FASTF score in the context of a database search is determined by the product of two terms: 1) the probability of obtaining an alignment score in a single pairwise alignment, and 2) the number of alternative alignments considered, a value which depends on the length of the sequences involved and the number of queries and database sequences that were compared. Term 1, the probability of a single pairwise alignment score $S$, $P(S \geq x)$, or $P_S$, can be calculated from the frequency of each amino acid and the scoring matrix (17–20). There are two sources of alternative alignments that contribute to the "search space" term 2 with FASTS and FASTF: (*a*) the alternative arrangements of peptides made possible by the length of the library sequence, and (*b*) the possibility that not all peptides will be aligned.

$K$ peptides of aggregate length $M$ can be aligned to a sequence of length $L$ in $N_A = ((L - M) + K)!/(L - M)!$ ways. If the $N_A$ alternative alignment scores are independently and identically distributed, the pairwise alignment probability becomes $P_{S \geq x|N_A} = 1 - \exp(-N_A P_S)$ when the alternative alignment search space term is considered (21). However, because of the FASTA heuristic strategy, not all of these possible positions will have been explored, thus reducing the actual alignment search space. Moreover, the $N_A$ different alignments do not generate truly independent scores because of local compositional effects and higher order sequence dependences. Together, this means that $N_A$ is too large; the correction overestimates the alignment search space and results in statistical estimates that are much too conservative.

Neither FASTS nor FASTF requires all of the query peptides to align; this allows for contaminating peptides or the simultaneous analysis of protein complexes. This adds another level of query complexity; for FASTS, $N_Q = 2^K - 1$ unique peptide selections may be obtained from a query containing $K$ peptides. This factor represents the maximal combinatorial search space explored during the sub-alignment joining stage. However, there are strong dependences between the scores obtainable by each of these combinations, so the correction is again conservative. For FASTF, the number of unique peptide selections and deconvolutions with $K$ peptides, each of length $M$ and having $K$ unique residues at each position, is as follows in Equation 1.

$$N_Q = K!^M \times \sum_{i=1}^{K} (K - i)!^{-M}/I! \qquad \text{(Eq. 1)}$$

The number represented by Equation 1 grows factorially with the number of peptides present in the query and exponentially with the length of the peptides. This second adjustment is again too large; it does not take into account the greatly reduced number of possible deconvolutions explored by the greedy residue consumption method used by FASTF.

The alignment search space-corrected probabilities $P_{S|N_A}$ are used to select optimal alignments during the initial search. These alignment probabilities are then scaled to reflect the apparent combinatorial search space size. FASTS and FASTF use the initial $P_{S|N_A}$ alignment probabilities to estimate an empirical combinatorial search space correction. The 95% of alignments with the highest probabilities most likely to be because of chance are fit to the equation, $\ln(P_{S|N_A}) = a\ln(P_O) + bP_O + c$, where $P_O$ is the observed frequency of alignments with probabilities better than or equal to $P_{S|N_A}$, and $a$, $b$, and $c$ are parameters to be estimated by multiple linear regression. This relationship fits the observed distribution of probabilities over its entire range (data not shown) and resembles mixed exponential decay; this is somewhat expected, as the FASTA algorithm is designed to find the best alignments of high quality while not spending any time optimizing an alignment of already low quality. After obtaining the parameter estimates $\hat{a}$, $\hat{b}$, and $\hat{c}$, the alignment probabilities $P_{S|N_A}$ are scaled to yield the final probability $P$ of each alignment. This value is used to report the statistical expectation estimate, $E = PN$, where $N$ is the number of sequences searched in the database.

*Database Searches with FASTS and FASTF*—Searches with FASTS and FASTF use a shallow scoring matrix with high information content (MDM20 (14) for protein databases or MDM10 for searches against DNA) because of the small amount of sequence content in each query. No gap penalties are used. A web interface to the programs is available at fasta.bioch.virginia.edu, and the source code is available as part of the FASTA source distribution, via FTP at ftp.virginia.edu/fasta/.

The probabilistic alignment strategy that improves search sensitivity is time consuming; a runtime option is available in which no alignment probabilities are calculated initially; only the raw alignment

score is used as the alignment optimality criterion. After the database has been searched, the top scoring 10% of sequences are realigned using the normal probability-driven alignment method and resorted for reporting. To obtain a distribution of alignment probabilities with which to perform the empirical scaling step described previously, additional randomly selected database sequences are realigned with a shuffled query, using probability-driven alignment. For all FASTF searches, and for those FASTS searches that recalculate compositional frequencies for each library sequence, using raw alignment scores as surrogates for the true alignment probabilities during the initial search of the database improves the run time by a factor of 10 or more, while not greatly effecting sensitivity (data not shown).

*Construction of the Test Database and Queries*—FASTS and FASTF performance was evaluated on a subset of proteins from the SwissProt v34 (22) database whose encoding DNA sequence was also available from GenBank™.[2] 111 protein families (defined by their PROSITE (23) and PFAM (24) annotations) from the test database were selected that satisfied the following criterion: from each family, a representative sequence could be selected that shared more than 50% sequence identity, over a region of more than 50 residues, with at least 15 other family members. Additionally, any family whose chosen representative sequence was able to identify a non-annotated sequence as statistically significant using the Smith-Waterman search algorithm was considered annotated incompletely and dropped from further usage.

Five equally spaced, non-overlapping 10-mer peptides were extracted from within the identified region of shared sequence identity in each representative sequence. These 111 queries, consisting of five peptides each, were used to generate successively smaller queries containing fewer peptides and of shorter length. This process was continued until all possible sets of nested queries were obtained consisting of between two and five peptides and of length between three and ten residues each. The described sequence databases and query datasets are available via anonymous FTP at ftp.virginia.edu/fasta/data/fastsf_data.tar.gz.

*Equivalence Number Calculation and the Sign Test*—Search performance was evaluated using equivalence numbers, a measure of the number of related sequences found in a search (25). If all related sequences are ranked higher than all unrelated sequences, the equivalence number is 0. For all other orderings, the equivalence number ranges between 1 and the size of the family of a given query. We use the non-parametric sign test statistic to assess any differences in performance indicated by the distribution of increases and decreases in equivalence numbers from independent queries.

*Comparison between FASTS and MS-Shotgun*—Fourteen MS/MS-derived FASTS queries from *Trypanosoma brucei* 20 S proteosomal proteins, published in Ref. 8, were used to search the National Center for Biotechnology Information (NCBI) non-redundant protein database (obtained October 11, 2001). We removed from consideration all hits against sequences from organisms in the *Kinetoplastida* taxonomic subtree (which includes *T. brucei*), as determined by the NCBI's Taxonomy database at ncbi.nlm.nih.gov/Taxonomy. FASTS *p* values are calculated from the reported expectation (*E*) values as $P = 1 - \exp(-E)$. Percent identities were obtained by aligning the full-length query sequence to the best related sequence identified by FASTS; alignment gaps were not counted in the percentage calculation. For spots where FASTS failed to report any related sequences, the corresponding full-length query sequences were used to search the *Kinetoplastida*-filtered non-redundant database with FASTA. MS-Shotgun *p* values are as reported in Ref. 8. Highest scoring unrelated library sequences were identified by searching the complete non-redundant protein database with the full-length candidate library sequence, from which no hits against proteosomal sequences were found with $E \leq 10^{-3}$.

[2] M.-Q. Huang and W. R. Pearson, manuscript in preparation.

## RESULTS

*Searching with FASTS and FASTF*—FASTS and FASTF searches with experimentally obtained sequence queries are shown below. FASTS queries use a modified FASTA format, shown below, with commas separating the query peptides.

```
>unknown FASTS query
QFLYEY,
PLVEET,
DETYA
```

This search against the SwissProt sequence database was performed on a 1-GHz Pentium III computer running the GNU/Linux OS and took ~10 s to complete. Below is shown the list of top scoring hits, identifying the protein as a serum albumin.

```
The best scores are:                    initE(94,006)
gi|113580|sp|P02770|ALBU_RAT     145 0.0038
gi|1351908|sp|P49064|ALBU_FELCA  145 0.004
gi|5915682|sp|P07724|ALBU_MOUSE  145 0.0042
gi|3121749|sp|035090|ALBU_MERUN  145 0.0043
gi|113576|sp|P02768|ALBU_HUMAN   144 0.0055
gi|1351909|sp|P49065|ALBU_RABIT  142 0.0085
...[10 more serum albumin hits with E()N<1.0]
```

The programs also output the calculated alignments for the top hits.

```
   query: ---QFLYEY---- ---PLVEET--- ---DETYA---
            :::::        :::::        ::::
ALBU_RAT: ..GTFLYEYSR.. ..QPLVEEPK.. ..VDETYVP..
                 360          410     520
```

FASTF uses the same format as FASTS, with random assignments of the residues identified at each cycle to a specific peptide. Thus, in the FASTF query shown below, the d, g, t, and l residues obtained in cycle 1 (m is the first residue, because the peptides were produced by cyanogen bromide cleavage) have been arbitrarily assigned to peptides 1 through 4. FASTF reads each column as a position (ignoring the vertical order of the residues within the column).

```
>unknown FASTF query
mdeaqwiyqraiv,
mgitkrseykpte,
mtldlglglftgq,
mlvexsvpxxxlk
```

Searching the NCBI non-redundant protein database (699,616 sequences) with this query took 80 s. The alignment shows that FASTF has identified the query as a ZIP-kinase (with an expectation of $2.7 \times 10^{-8}$) and deconvoluted the input sequence while preserving the positional composition defined by the query. In this case, however, the alignment only involves three of the four peptides.

```
     query: ---MGEELGSGQFAIV--- ---MLLDKRVPXRPLQ---
              :::::::::::::       ::::: ::.
ZIP-kinase: ..EMGEELGSGQFAIVR.. ..IMLLDKNVPNPRIKL..
                   20              150       160
     query: -------------------MTIAQSLEYXXTK-----
                               :::::::::.. :
ZIP-kinase: SELAKDFIRRLLVKDPKRRMTIAQSLEHSWIKAI...
                  260       270       280
```

*Accuracy of Estimates of Statistical Significance*—Since the introduction of the BLAST program for rapid sequence simi-

larity searches (26), most widely used sequence comparison programs provide an estimate of how frequently an alignment score is expected by chance. If the statistical estimates are accurate, then an unrelated sequences should have alignment scores with an expectation $E$ of 0.01 in about 1% of independent searches, expectations of $E \leq 0.001$ should be seen one time in 1000, etc. If the highest scoring unrelated sequence obtains $E \leq 0.1$ only once in 1000 searches, the estimates are too conservative, and related sequences are likely to be missed as false negatives (type II errors). Conversely, if an unrelated sequence receives an $E \leq 0.01$ in every search, many false positive (type I) errors are likely to occur. Thus, in evaluating the performance of a sequence comparison strategy, it is important to examine the accuracy of its statistical estimates.

To evaluate the accuracy of FASTS and FASTF statistical estimates, we used our test queries to search annotated protein and DNA databases, examining expectation estimates of the highest scoring unrelated sequence from each search (Fig. 2). Using FASTF, independent queries sharing the same peptide number and length may exhibit modest type I statistical error (Fig. 2*A*); most FASTS estimates are very reliable. FASTF estimates are less accurate, often lower by factors of 20–50. Statistical inaccuracy depends on both the length and number of peptides, and translated sequence comparisons (TFASTS and TFASTF) provide less accurate statistical estimates (Fig. 2*B*). Error increases with total query content and is generally about 10-fold worse with searches against DNA databases, an effect that is also seen in standard translated DNA search algorithms (27).

Fig. 2*B* provides a guide for choosing conservative expectation thresholds appropriate to algorithm and query content; for example, to obtain a false positive error rate of 0.01, for the average FASTS query consisting of three or four peptides, each having a length of four or five residues, a conservative expectation threshold would be $10^{-3}$ and another 10-fold smaller for a TFASTS search. For an average FASTF query content of three peptides of length 10, a threshold of $10^{-4}$ would be appropriate and another 10-fold smaller when using TFASTF.

*Evaluation of Alignment Probability as Optimality Criterion*—Most sequence alignment algorithms, including some designed for use with MS/MS-derived sequence (28), maximize the sum total similarity score of an alignment, rather than minimizing the probability of obtaining the alignment by chance (29–31). In searches with multiple peptides, however, whereas any single peptide involved in an alignment with a higher similarity score will result in a lower probability of the score $P_S$, the addition of a second peptide to an existing alignment may not produce a more statistically significant alignment under the statistical model we describe; the additional peptide increases the alignment search space adjustment $N_A$. To take this potential penalty into account, FASTS and FASTF use the adjusted probabilities $P_{S|N_A}$ when joining

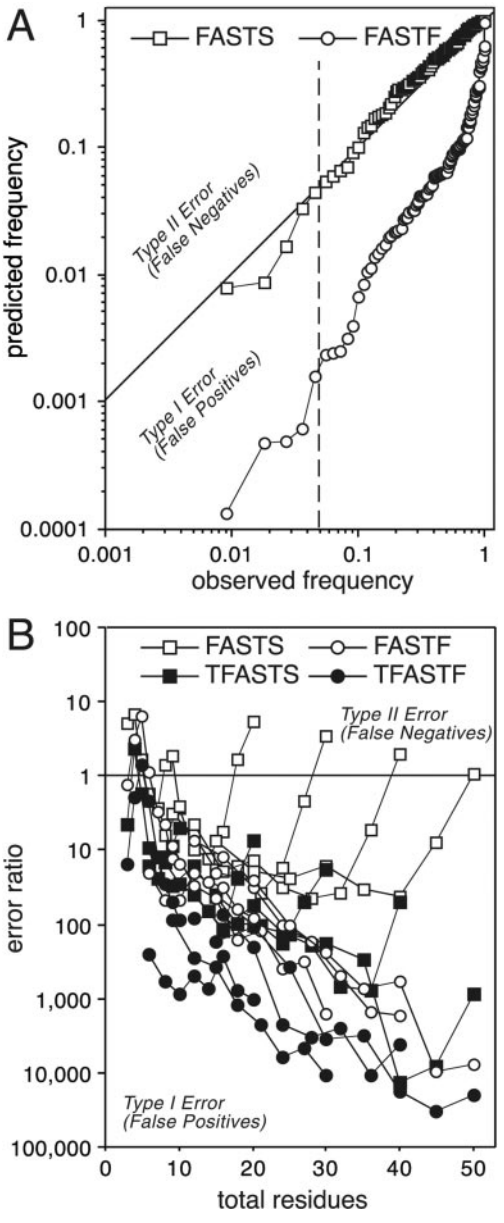


FIG. 2. **Accuracy of FASTS and FASTF statistical estimates.** *A*, each of 111 independent queries with three peptides of length 8 were compared against an annotated protein sequence database. The predicted frequency of the highest scoring unrelated sequence alignment is plotted on the ordinate against the observed frequency of the alignment probability. Ideally, the predicted and observed frequencies will be identical, as indicated by the *diagonal line*. When the predicted frequencies are too low (*points below* the *diagonal line*), the statistical significance of a match will be overestimated. The *dashed vertical line* indicates the 95th percentile of searches, used to evaluate accuracy in *B*. *B*, for each of the 111 independent test queries, replicate searches were performed with varying sequence content (total residues). Peptide number varied from 2 to 5; peptide length varied from 3 to 10. The error ratio observed in the predicted frequency of the highest scoring unrelated sequence at the 95th percentile is shown for each combination of peptide length and number and for protein *versus* translated DNA searches using TFASTS and TFASTF. *Data points* that are derived from query sets sharing the same number of peptides are connected by *solid lines*.

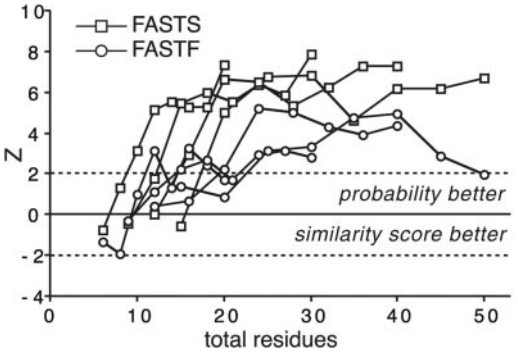


FIG. 3. **Probabilistic alignments are more sensitive than score-based alignments.** The graph represents the ability of FASTS and FASTF to distinguish related from unrelated sequences when using alignment probabilities or similarity scores measured by the equivalence number (see "Experimental Procedures"). Shown are sign test Z values from comparisons with varying total sequence content (from 2 to 5 fragments of length 3–10), using all 111 test queries against the curated protein database. *Data points* from queries with the same number of peptides are connected by *lines*. Positive sign test Z values indicate better performance by probabilistic alignments. Differences in performance with $Z > 2.0$ or $Z < -2.0$ are statistically significant at the $p = 0.05$ threshold.

subalignments. This optimality criterion requires multiple-peptide alignments to be composed of higher scoring subalignments, excluding low (but positive-) scoring subalignments that would otherwise worsen the overall alignment quality (an artifact often termed the "mosaic effect" (32)).

With both FASTS and FASTF, we find that, except for queries with vanishingly small sequence content, probabilistic alignments provide better discriminatory power than subalignment joining based on similarity scores alone (Fig. 3). We also measured the direct effect of probabilistic scoring on sensitivity at a conservative expectation threshold and find that as many as 30% more related sequences at 80–100% identity could be identified with probabilistic alignment optimization; at 50–70% identity, however, a more modest 10% improvement is seen.

*Sensitivity of the Algorithms*—We also examined the sensitivity of FASTS and FASTF at specific evolutionary distances, as measured by average percent identity (Fig. 4). Both algorithms readily identify statistically significant alignments between distantly related sequences. Even at a target of ~70% identity, FASTS can identify over 50% of the related sequences in the database using four or five peptides of length 5. Generally, FASTF searches require more sequence content to perform similarly, requiring three or four peptides of length 9 to achieve the same sensitivity as FASTS in the previous example. These length and peptide number requirements are well within the bounds of achievable sequence data for each target experiment.

Against the equivalent DNA test database, TFASTS and TFASTF achieve sensitivity levels almost as good as those of the protein sequence-based searches (data not shown). This is expected, as our DNA test database is composed of primarily cDNA nucleotide sequences, and so the increase in the total database search space (~6-fold greater) was not enough to drastically reduce the number of statistically significant homologous sequences. Thus, FASTS and FASTF can be used to search expressed sequence tags or unfinished, unannotated genome databases.

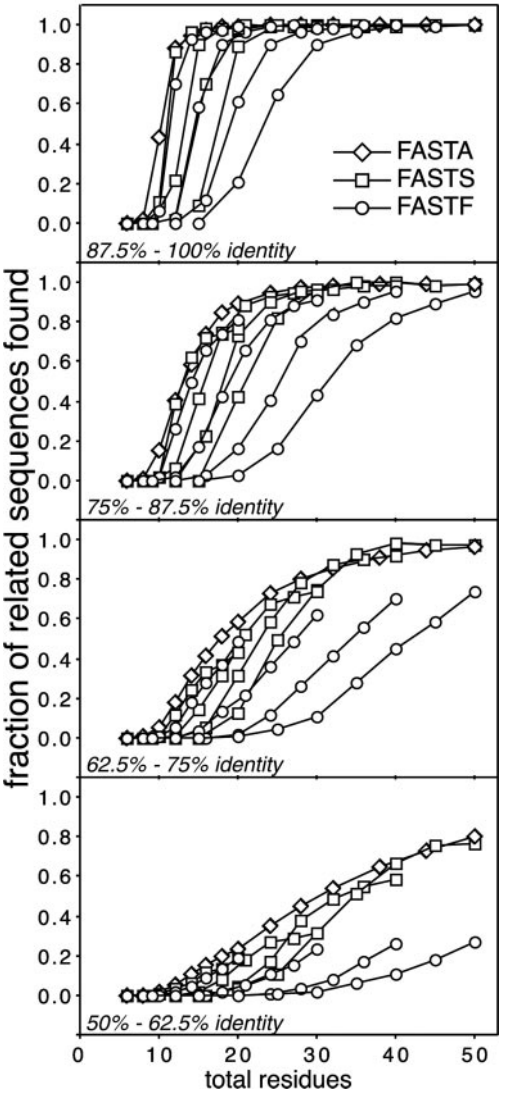


FIG. 4. **FASTS and FASTF identify statistically significant alignments between related family members with modest amounts of query content.** The ability of each algorithm to identify statistically significant family members at varying evolutionary distance using queries of varying sequence content is shown. Plotted on the *y* axis is the mean fraction of family members at each percent identity range that were identified by the algorithms with expectations better than $10^{-4}$, using default parameters. *Data points* that are derived from query sets sharing the same number of peptides are connected by *solid lines*. *Data points* labeled as FASTA are derived from the geometric means of 10 separate FASTA searches (using an unmodified MDM20 scoring matrix and default gap penalties) with single peptides extracted randomly from within the conserved domain from the reference sequence of each family.

Searches with a smaller number of longer peptides are

TABLE II
*FASTS and MS-Shotgun identification of T. brucei 20 S proteosomal proteins*

| Spot | Peptides | Length | % Identity | Best MS-Shotgun *P* | | Best FASTS *P* | |
|---|---|---|---|---|---|---|---|
| | | | | Related | Unrelated | Related | Unrelated |
| 17 | 5 | 54 | 76.5 | $8.0 \times 10^{-12}$ | $8.5 \times 10^{-3}$ | $9.1 \times 10^{-28}$ | 1.00 |
| 4 | 7 | 79 | 50.2 | $3.8 \times 10^{-24}$ | $3.8 \times 10^{-4}$ | $1.1 \times 10^{-21}$ | 0.95 |
| 11 | 6 | 77 | 57.1 | $2.5 \times 10^{-16}$ | $7.5 \times 10^{-4}$ | $2.2 \times 10^{-20}$ | 0.89 |
| 1 | 3 | 46 | 55.6 | $7.9 \times 10^{-6}$ | $1.7 \times 10^{-2}$ | $3.5 \times 10^{-14}$ | 0.98 |
| 7 | 4 | 53 | 55.3 | $2.5 \times 10^{-11}$ | $7.0 \times 10^{-5}$ | $1.2 \times 10^{-13}$ | 0.75 |
| 15[a] | 2 | 22 | 46.8 | $2.6 \times 10^{-2}$ | 1.00 | $7.8 \times 10^{-9}$ | 0.97 |
| 6 | 5 | 55 | 40.7 | $4.9 \times 10^{-8}$ | 0.95 | $1.7 \times 10^{-6}$ | 0.78 |
| 5[a] | 7 | 67 | 41.5 | $8.8 \times 10^{-3}$ | $7.0 \times 10^{-6}$ | $5.0 \times 10^{-6}$ | 0.22 |
| 2[a] | 5 | 49 | 47.1 | $3.4 \times 10^{-4}$ | $4.0 \times 10^{-3}$ | $2.0 \times 10^{-3}$ | 0.99 |
| 12[a] | 4 | 35 | 44.6 | $1.3 \times 10^{-2}$ | $1.3 \times 10^{-2}$ | $2.5 \times 10^{-3}$ | 0.98 |
| 3[a] | 4 | 48 | 42.0 | $6.4 \times 10^{-2}$ | $3.6 \times 10^{-2}$ | $4.9 \times 10^{-2}$ | 0.99 |
| 9[a] | 4 | 39 | 44.3[b] | $3.9 \times 10^{-2}$ | $1.8 \times 10^{-2}$ | 1.00 | 1.00 |
| 13[a] | 3 | 29 | 64.1[b] | 1.00 | 0.25 | 1.00 | 1.00 |
| 8[a] | 3 | 27 | 39.8[b] | 1.00 | 0.61 | 1.00 | 0.80 |

[a] Queries that were difficult to identify by MS-Shotgun ($P_{rel}/P_{unrel} > 10^{-2}$).
[b] Identified and measured by FASTA.

more sensitive, particularly at greater evolutionary distance. Reduced sensitivity with more peptides largely reflects an increase in the theoretical $N_Q$ term and associate increase in $P_{S|N_A}$ scaling. With FASTS queries, the penalty for additional peptides is relatively small and is easily offset by the gain in total information content afforded by the extra residues. Thus, FASTS sensitivity nearly always increases with additional peptides of similar length (Fig. 4). When data from an MS/MS experiment fails to find a significant hit, sequence data obtainable from interpretation of additional MS fragment spectra should improve sensitivity. In contrast, FASTF queries suffer large penalties with the addition of peptides, made even worse as the peptide length increases (Fig. 4). Unlike MS/MS experiments, however, in a mixed Edman degradation experiment there is little control over the number of peptides from which sequence is obtained. Luckily, this effect is mitigated by the ability of Edman sequencing to generate longer peptide sequences that overcome the combinatorial penalty.

*Comparison of FASTS to Alternative Methods*—Both FASTA (CIDentify (33)) and BLAST (MS-BLAST (28), MS-Shotgun (8)) have been employed by previous methods to search databases with MS/MS-derived sequence data. These earlier methods use various forms of congruency analysis to identify database sequences that hit with the highest scores and most often against the peptide sequences in each query. Of the three, only MS-Shotgun attempts to align all of the query peptides simultaneously (by repeating gapped-BLAST searches with all possible permutations of the peptide order of a query) and to assign statistical significance to the results. Therefore, we compared FASTS with respect to MS-Shotgun by repeating the analysis performed in Ref. 8. Fourteen experimentally obtained MS/MS peptide sequence queries from the 20 S proteasome subunit of *T. brucei* were used to search the National Center for Biotechnology Information non-redundant database of protein sequences, with all taxonomically adjacent *Kinetoplastida* sequences removed.

Although FASTS and MS-Shotgun performances are similar (Table II), FASTS statistical estimates are considerably more accurate than those produced by MS-Shotgun. The highest scoring unrelated sequences in the FASTS searches had *p* values ranging from 0.22 to 1.0; MS-Shotgun *p* values ranged from $10^{-5}$ to 1.0. This wide range of *p* values for unrelated sequences confounds attempts to identify unambiguously homologous database sequences. The importance of accurate statistical estimates can be seen clearly in the MS-Shotgun results for spot 5, where a significant alignment to a related sequence has a worse probability than that of an unrelated sequence; FASTS has no such difficulty. Although in Ref. 8 spots 2, 12, and 15 were all determined to be identifiable, the highest scoring homologs had *p* values worse than $10^{-4}$, and the related sequence had less than 100-fold differences in probability between related and unrelated sequences.

Three of the four unidentified queries (spots 3, 8, and 9) cannot be identified, because their closest homologs are too distant; the queries share 40–50% identity with their nearest homologues in the database. Spot 13 does have a nearest homologue that shares 65% identity overall, but the query peptides originate from poorly conserved portions of the sequence that shares less than 50% local identity. These results show that the 50% identity threshold for robust detection observed in Fig. 4 is consistent with the performance of real data against much larger databases.

## DISCUSSION

FASTS is designed to interpret *de novo* MS/MS data from organisms that lack comprehensive proteome sequence data, *e.g.* mammals other than humans and mice or plants other than *Arabidopsis*. Based on the results in Fig. 4, we expect FASTS to reliably identify more than 80% of sequences that share 65% identity if 30 amino acids of *de novo* sequence

data are available; for proteins with average divergence rates (10–30% per 100 million years) 65% identity would include proteins that diverged in the past 150–500 million years.

Although FASTS takes into account single residue isobars (I/L and Q/K), it does not correct for other sources of sequence error from spectral misinterpretation (*e.g.* dipeptide isobars, reversed sequence order). If such sources of error are likely in an experiment, additional peptide sequences reflecting these alternatives may be added to the FASTS query. These additional peptide sequences will incur a small penalty for the additional combinatorial complexity; for a query with five peptides, adding five reversed peptide sequences will increase search time by 2-fold and decrease the statistical significance of a match by $2^5 = 32$-fold. Because the addition of each additional peptide decreases significance by a factor of two, the inclusion of all possible sequence variants (however unlikely) is unadvisable. This robustness of FASTS to inclusion of peptides that may not be involved in any specific protein alignment makes it an ideal tool to simultaneously identify multiple proteins from mixtures (34); we have simultaneously identified multiple unrelated proteins in several experiments. Future versions of FASTS may be designed to analyze peptide data from more complex mixtures.

No algorithmic equivalent to FASTF currently exists. Mutation-sensitive pattern or motif search algorithms could be used to search a database with mixed Edman degradation-derived sequence data, but all matching sequences would still require further processing to determine which alignment assemblies satisfy the compositional requirements of the query, akin to the subalignment joining performed by FASTF. We are currently exploring methods to generate optimal FASTF alignments for display, correcting those mistakes made by the greedy alignment heuristic. We will also then evaluate whether taking the time to calculate optimal alignments during the database search has any measurable effect on sensitivity.

The probabilistic optimality criterion in FASTS improves search sensitivity over methods based on total similarity score alone (see Fig. 3, *e.g.* CIDentify and MS-BLAST). In a concrete example, a query consisting of five peptides of total length 35 from GBB3_RAT (guanine nucleotide-binding protein beta, subunit 3 from rat) achieves a nearly complete alignment against various coronavirus glycoproteins including VGL2_CVBV.

```
>VGL2_CVBV E2 Glycoprotein precursor (1363 aa)
 init: 170 %id: 62.857% E(): 2.9

 GBB3_RAT ---ASCRLFD-----NVKVSRE---  ---GVLSGHD--
             ::: :.       :: :::         ::    ::
 VGL2_CVBV..AASCQLYYNLPAANVSVSRFN..  ..VGVFTHHDV..
                430        440             470

 GBB3_RAT ---LAVSPDY---   ---SQDGKLI---
             ::.  ::         ::  :::
 VGL2_CVBV..GLAIKSDYC..  ..LSQNQKLIA..
                570    1010      1020
```

Inspection of the alignment suggests that this is a potentially homologous match; it has a very high similarity score (init = 170). However, the statistical exception provided by FASTS is only 2.9. Lower scoring (init = 153) yet more significant alignments ($E < 8.8 \times 10^{-5}$) occur with true homologs of GBB3_RAT. Probabilistic scoring, combined with accurate statistical estimates, makes FASTS a clear choice over score-based alternatives.

FASTS and FASTF achieve high sensitivity by maximizing the search potential of queries, with high information content scoring matrices, ungapped global peptide alignments, and a stringent probabilistic criterion for alignment optimality. Sensitivity can be improved by reducing the set of library sequences examined, for example by filtering the database by approximate molecular weight or isoelectric point (pI) ranges or by selecting a taxonomic subset of the data (mammals, plants, fungi). These options are all available within the FASTA search package.

*Acknowledgments*—We thank Stephen Altschul for initially suggesting the method used to assess statistical significance of each peptide subalignment. We thank Ming-Qian Huang for work constructing our curated SwissProt subset database and Raphaël Clifford for combinatorics expertise. Ken Mitchelhill provided the experimentally obtained FASTS query examples.

* 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 Material.

§ Supported by Grant T32AI07046 from the National Institutes of Health.

‖ Supported by Grants HL19242-24 and DK52378-04 from the National Institutes of Health.

‡‡ Supported in part by Grant LM04969 from the National Library of Medicine, with additional support from the Compaq Computer Corporation. To whom correspondence should be addressed. Tel.: 434-924-2818; Fax: 434-924-5069; E-mail: wrp@virginia.edu.

REFERENCES

1. Eng, J. K., McCormack, A. L., and Yates, J. R., III (1994) An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J. Am. Soc. Mass Spectrom.* **5,** 976–989
2. Pevzner, P. A., Dancik, V., and Tang, C. L. (2000) Mutation-tolerant protein identification by mass spectrometry. *J. Comput. Biol.* **7,** 777–787
3. Pevzner, P. A., Dancik, V., Mulyukov, Z., and Tang, C. L. (2001) Efficiency of mutation-tolerant database search with tandem mass spectra. *Genome Res.* **11,** 290–299
4. McLafferty, F. W., Fridriksson, E. K., Horn, D. M., Lewis, M. A., and Zubarev, R. A. (1999) Biomolecular mass spectrometry. *Science* **284,** 1289–1290
5. Dancik, V., Addona, T. A., Clauser, K. R., Vath, J. E., and Pevzner, P. A. (1999) *De novo* peptide sequencing via tandem mass spectroscopy. *J. Comput. Biol.* **6,** 327–342
6. Chen, T., Kao, M., Tepel, M., Rush, J., and Church, G. M. (2000) in *Proceedings of the ACM-SIAM Symposium on Discrete Algorithms, San Francisco, 2000*, pp. 389–398, ACM Press, New York
7. Mann, M., and Wilm, M. (1994) Error-tolerant identification of peptides in sequence databases by peptide sequence tags. *Anal. Chem.* **66,** 4390–4399
8. Huang, L., Jacob, R. J., Pegg, S. C.-H., Baldwin, M. A., Wang, C. C., Burlingame, A. L., and Babbitt, P. C. (2001) Functional assignment of the 20 S proteasome from *Trypanosoma brucei* using mass spectrometry and new bioinformatics approaches. *J. Biol. Chem.* **276,** 28327–28339
9. Altschul, S. F. (1991) Amino acid substitution matrices from an information

theoretic perspective. *J. Mol. Biol.* **219,** 555–565
10. Altschul, S. F., Boguski, M. S., Gish, W., and Wootton, J. C. (1994) Issues in searching molecular sequence databases. *Nat. Genet.* **6,** 119–129
11. Damer, C. K., Partridge, J., Pearson, W. R., and Haystead, T. A. J. (1998) Rapid identification of protein phosphatase 1-binding proteins by mixed peptide sequencing and database searching. Characterization of a novel holoenzymic form of protein phosphatase 1. *J. Biol. Chem.* **273,** 24396–24405
12. Pearson, W. R., and Lipman, D. J. (1988) Improved tools for biological sequence comparison. *Proc. Natl. Acad. Sci. U. S. A.* **85,** 2444–2448
13. Schwartz, R. M., and Dayhoff, M. (1978) in *Atlas of Protein Sequence and Structure* (Dayhoff, M., ed) Vol. 5, Suppl. 3, pp. 353–358, National Biomedical Research Foundation, Silver Spring, MD
14. Jones, D. T., Taylor, W. R., and Thornton, J. M. (1992) The rapid generation of mutation data matrices from protein sequences. *Comput. Appl. Biosci.* **8,** 275–282
15. Karlin, S., and Altschul, S. F. (1990) Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proc. Natl. Acad. Sci. U. S. A.* **87,** 2264–2268
16. Altschul, S. F., and Gish, W. (1996) Local alignment statistics. *Methods Enzymol.* **266,** 460–480
17. McLachlan, A. D. (1983) Analysis of gene duplication repeats in the myosin rod. *J. Mol. Biol.* **169,** 15–30
18. Staden, R. (1989) Methods for calculating the probabilities of finding patterns in sequences. *Comput. Appl. Biosci.* **5,** 89–96
19. Tatusov, R. L., Altschul, S. F., and Koonin, E. V. (1994) Detection of conserved segments in proteins: iterative scanning of sequence databases with alignment blocks. *Proc. Natl. Acad. Sci. U. S. A.* **91,** 12091–12095
20. Bailey, T. L., and Gribskov, M. (1998) Combining evidence using *p*-values: application to sequence homology searches. *Bioinformatics* **14,** 48–54
21. Arratia, R., Gordan, L., and Waterman, M. S. (1990) The Erdos-Renyi law in distribution, for coin tossing and sequence matching. *Ann. Stat.* **18,** 539–570
22. Bairoch, A., and Boechmann, B. (1991) The SWISSPROT protein sequence data bank. *Nucleic Acids Res.* **19,** (suppl.) 2247–2249
23. Bairoch, A. (1991) PROSITE: a dictionary of sites and patterns in proteins. *Nuc. Acids Res.* **19,** (suppl.) 2241–2245
24. Sonnhammer, E. L., Eddy, S. R., and Durbin, R. (1997) Pfam: a comprehensive database of protein domain families based on seed alignments. *Proteins* **28,** 405–420
25. Pearson, W. R. (1995) Comparison of methods for searching protein sequence databases. *Protein Sci.* **4,** 1145–1160
26. Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990) Basic local alignment search tool. *J. Mol. Biol.* **215,** 403–410
27. Pearson, W. R., Wood, T., Zhang, Z., and Miller, W. (1997) Comparison of DNA sequences with protein sequences. *Genomics* **46,** 24–36
28. Shevchenko, A., Sunyaev, S., Loboda, 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
29. Altschul, S. F., and Erickson, B. W. (1986) A nonlinear measure of subalignment similarity and its significance levels. *Bull. Math. Biol.* **48,** 617–632
30. Althschul, S. F., and Erickson, B. W. (1986) Locally optimal subalignments using nonlinear similarity functions. *Bull. Math. Biol.* **48,** 633–660
31. Altschul, S. F., and Erickson, B. W. (1988) Significance levels for biological sequence comparison using non-linear similarity functions. *Bull. Math. Biol.* **50,** 77–92
32. Arslan, A. N., Egecioglu, O., and Pevzner, P. A. (2001) A new approach to sequence comparison: normalized sequence alignment. *Bioinformatics* **17,** 327–337
33. Taylor, J. A., and Johnson, R. S. (1997) Sequence database searches via *de novo* peptide sequencing by tandem mass spectrometry. *Rapid Commun. Mass Spectrom.* **11,** 1067–1075
34. Link, A. J., Eng, J., Schieltz, D. M., Carmack, E., Mize, G. J., Morris, D. R., Garvik, B. M., and Yates, J. R., III (1999) Direct analysis of protein complexes using mass spectrometry. *Nature Biotechnol.* **17,** 672–682