# Probability-Based Protein Identification for Post-Translational Modifications and Amino Acid Variants Using Peptide Mass Fingerprint Data

Tong WW, McComb ME, Perlman DH, Huang H, O'Connor PB, Costello CE, and Weng Z

# Overview

- A web server that interprets peptide mass fingerprinting data by searching the spectrum against protein sequence databases using log likelihood ratio as scoring function.
- Log likelihood ratio is the best statistical model to distinguish correctly assigned peptides from incorrect assignments.
- Matching peaks with self-adjusting mass tolerance offers more flexibility and accuracy than the traditional mass window.
- Seamless pipeline for further processing of the result (internal calibration and PTM searches).
- BUPID outperforms conventional database search algorithms in sensitivity and specificity.

## Introduction

- A major goal of proteomics is to identify and characterize all proteins expressed in cells in various conditions. Mass spectrometry (MS) has become popular for identification of proteins in high-throughput proteomics research. Peptide mass fingerprint (PMF) with database search is a standard way to analyze high-throughput MS data of complex protein samples. This method compares the PMF spectrum against those expected for all possible proteins obtained from a sequence database. Each database protein is then assigned a probability score that reflects various aspects of the fit between spectrum and peptide. These scores help discriminate correct and incorrect protein assignments to the spectrum and therefore identify proteins in the sample.

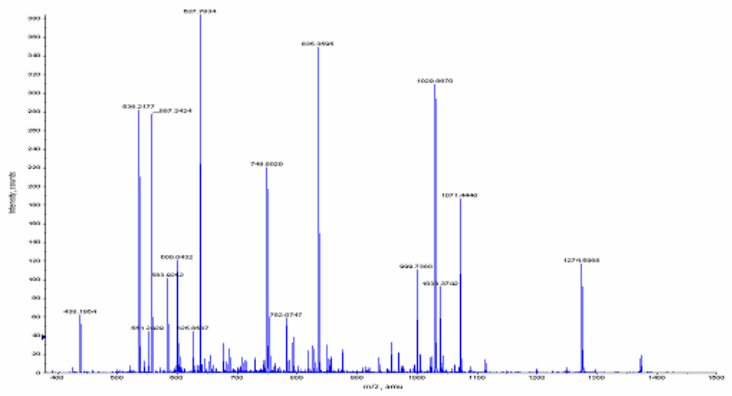
# BUPID

http://zlab.bu.edu/Amemee

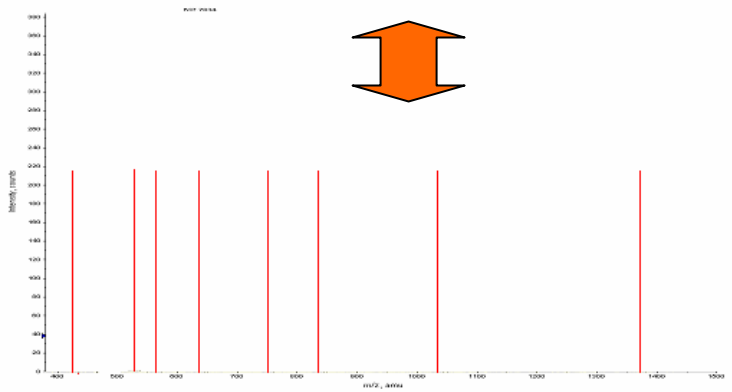
- We report a new method of database searching using MS data. The algorithm utilizes a log likelihood score to discriminate correctly assigned peaks from incorrectly assigned ones. The accumulative log likelihood score of all peaks that match with the protein represent the significance of the correlation between the protein and the spectrum. The raw score is normalized according to the size and mass of the protein in order to compare with other proteins against the same standard. The method was proven to have the best sensitivity and specificity in identifying proteins in a mixture. Currently BUPID only search for unmodified proteins. PTM search is under development and will be available to the public shortly.

# The Principle of Protein Identification by Peptide Mass Fingerprinting

VLSPADKTNIKTTWDKLGGHAGEYGGEALER
TFMSFPTTKTYFPHFDLSPGSAQVKAHGKKVA
DALTTAVAHMDDLPKALSALSDLHAYKLRVD
PVNFKLLSHCLLVTLACHHPAEFTPAVHASLD
KFFSTVSTVLTSKYR



| | |
|---|---|
| VLSPADK | 728.83436 |
| TNIK | 474.55196 |
| TTWDK | 649.6929 |
| LGGHAGEYGGEALER | 1515.58392 |
| TFMSFPTTK | 1059.23686 |
| TYFPHFDLSPGSAQVK | 1793.972 |
| AHGK | 411.45634 |
| K | 146.18764 |
| VADALTTAVAHMDDLPK | 1767.99884 |
| ALSALSDLHAYK | 1288.44932 |
| LR | 287.3588 |
| VDPVNFK | 817.92912 |
| LLSHCLLVTLACHHPAEFTPAVHASLDK | 3024.52148 |
| FFSTVSTVLTSK | 1316.4991 |
| YR | 337.37442 |



All database searching algorithms, although differ in scoring functions, share a similar approach. First, theoretical peptide mass values are created by applying simulative cleavage to sequences in the database. Then a set of peaks (matches) in the experimental spectrum are picked out and compared with calculated peptide mass values of the protein. Keep in mind that identifying matched peaks is as important as assessing their similarities.

# Incentive to Use and Not to Use Peptide Mass Fingerprinting

- PMF is faster to acquire
- PMF is more sensitive
- PMF requires full (and correct) protein sequences in the databases

# BUPID Uses Log Likelihood Ratio as the Target Function

$$score = \ln \frac{likelihood\ H_A}{likelihood\ H_0}$$

$$= \ln \frac{likelihood\left(spectrum\ data \mid predicted\ \ protein\right)}{likelihood\left(spectrum\ data \mid random\ background\right)}$$

Spectrum data: All possible *combination* of peaks.
Predicted protein: A protein considered to be in the sample.
Background:
- All proteins in the mixture.
- All proteins in the database.
- All proteins in the universe.

Log likelihood ratio is considered the most powerful method to distinguish two hypotheses. In the database search problem, the null hypothesis is that a certain set of peaks in the spectrum is generated by random background noise. The alternative hypothesis is that the set of peaks is generated by the peptides corresponding to the predicted protein.

# Calculation for the Likelihood of the Alternative Hypothesis

$$likelihood(spectrum\ data \mid prediction)$$

$$= \prod_{all\ peaks\ in\ the\ spectrum} [likelihood(peak \mid peptides)]$$

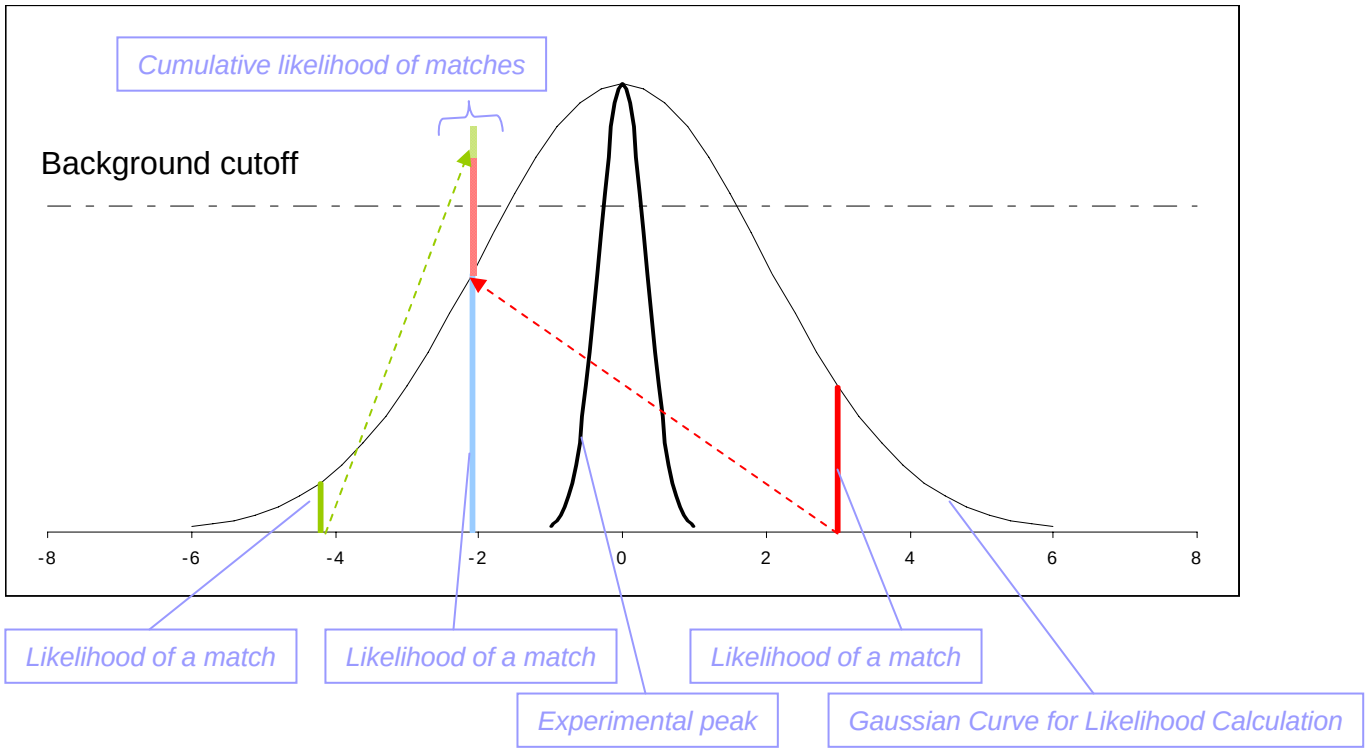
$$= \prod_{all\ peaks\ in\ the\ spectrum} \left[ \prod_{all\ peptide\ from\ digesting\ the\ target\ protein} [likelihood(peak \mid peptide)] \right]$$

Define that the peak is a match if likelihood of the match > background
(log likelihood ratio > 0)

and a miss-match if likelihood of the match <= background
(log likelihood ratio <= 0)

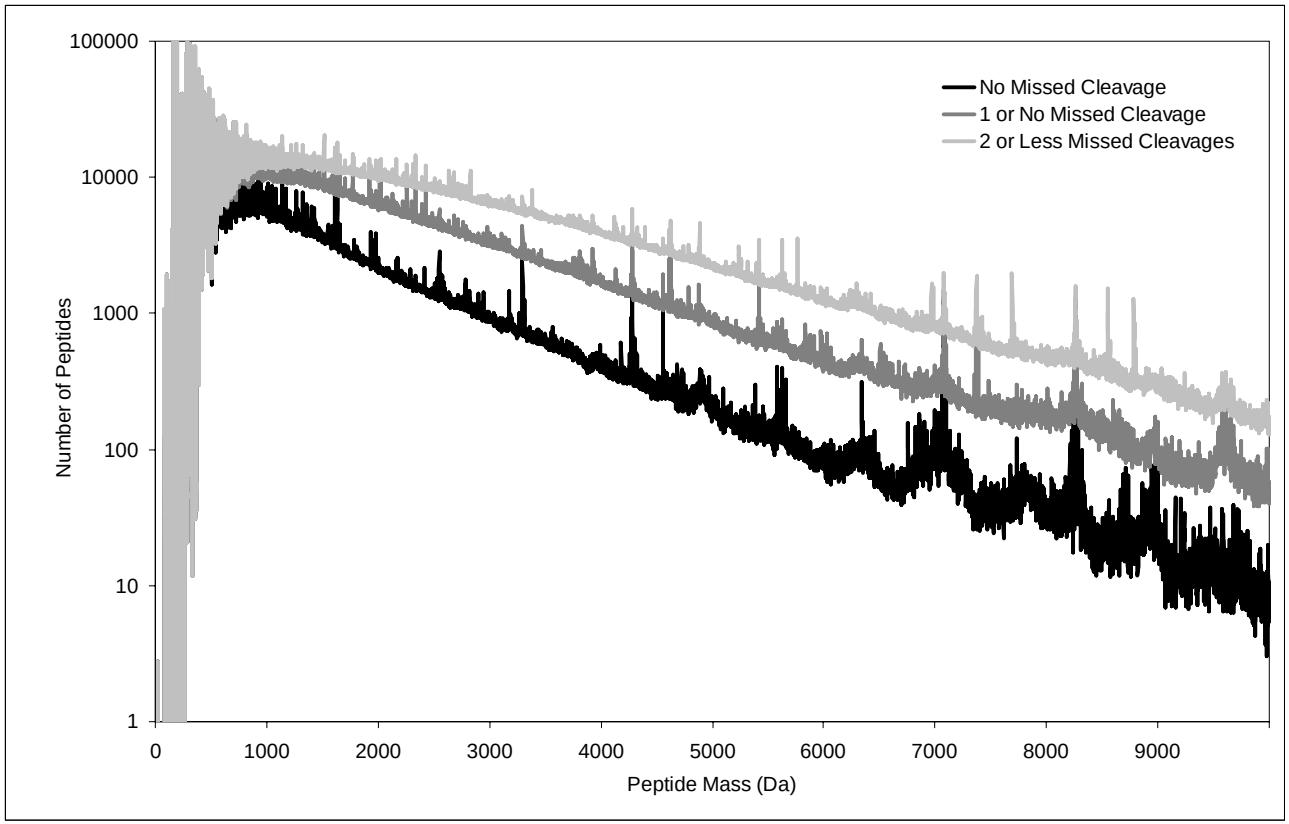
Automatically, the log likelihood of the alternative hypothesis is maximized if and only if all matched peaks and no miss-matched peaks are included in the calculation.

# Calculation for Likelihood of Matches




BUPID assumes that the experimental error follows the Gaussian Distribution. Thus the likelihood of a match is the z score of the mass difference. However if an experimental peak is adjacent to several theoretical peaks (peptides from the same protein or different proteins), chances are it may not be the result of the closest one. Thus it is necessary to calculate the cumulative likelihood of the experimental peak given all possible peptides exist in the sample. The bottom line is that any peak can be the result of a peptide with any mass. An far-fetched match still stands a chance in the nature, although much more impossible than a hit right in the bull's eye.

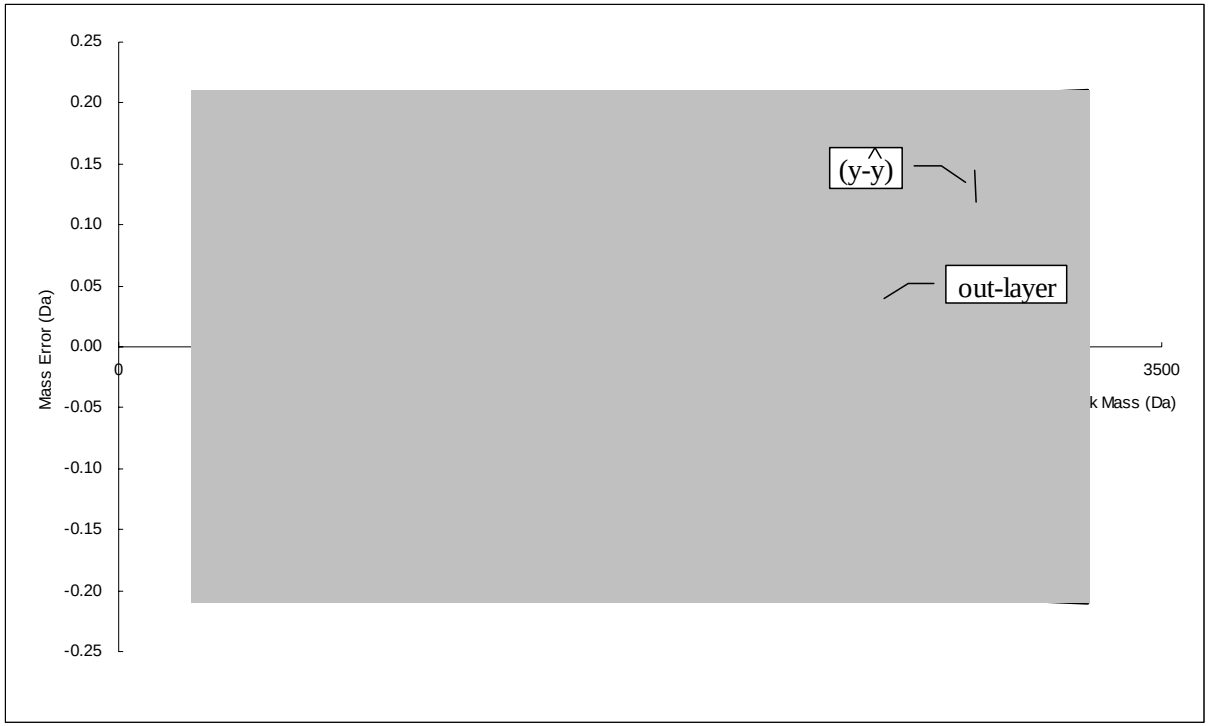
# Background probability




The background probability is generated by digesting all proteins the database. Assume that restriction sites appear randomly in the protein sequence with probability *P*. Length of peptides generated from the digestion would, therefore, be proportional to the exponential of (1-*P)*. Peptide mass roughly follows the same distribution, as shown in the plot. Notice that miss cleavages generate more peptides and have larger peptide mass in average.

Lines in red/yellow, green and blue each correspond to no miss cleavage, one miss cleavage, and one or two miss cleavages.

# Mass difference between peaks and their corresponding peptides




- Peptide mass

  Due to the non-uniform background, the log likelihood ratio of a match is affected by both the difference and the absolute mass of the peak-peptide pair. The threshold (the point where the likelihood that a peak is generated by the specific peptide is equal to the likelihood that the peak is generated by a random background noise) rises as peptide mass becomes larger.

- Number of peaks in the spectrum

  The more peaks in the spectrum, the easier for a protein to match and get a high score. Therefore the background probability is adjusted according to the number of peaks in the spectrum. This serves as a internal quality-control scheme.

# Correction for larger and/or heavier proteins

The matching score of a protein is the sum of the log likelihood ratios of all matched peptides, adjusted for the size and mass of the protein.

- Size of the protein

  Proteins with more restriction sites in their sequences generate more peptides after the simulated digestion. They are in advantage against small proteins to match more experimental peaks in the spectrum. This tendency has to be adjusted before all scores could be compared on the same basis.

- Mass of the protein

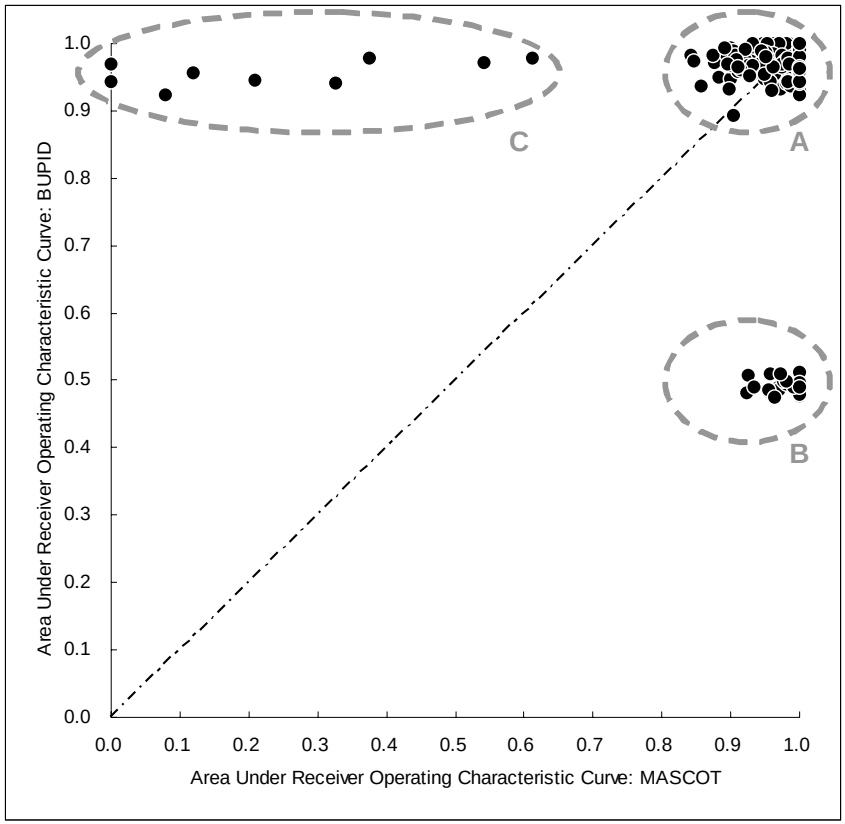
  Once again, peaks with smaller mass are easier to match with peptides because there are so many of them. The fact that light-weighted peptides are more abundant shifts the peptide mass tolerance and favors large proteins in scoring. This is also normalized before the search results are ranked.

# Result comparison: Human Blood Sample

| | MASCOT SEARCH RESULT | | | | | BUPID SEARCH RESULT | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Sample | HBA | HBB | HBG1 | HBG2 | HBD | HBA | HBB | HBG1 | HBG2 | HBD |
| A12 | 2 | 1 | 6 | 4 | --- | 2 | 1 | 7 | 3 | 4 |
| A02 | 2 | 1 | --- | --- | --- | 2 | 1 | --- | --- | 3 |
| A17 | 2 | 1 | --- | --- | 57 | 2 | 1 | --- | 15 | 3 |
| B1 | 2 | 1 | --- | --- | --- | 1 | 2 | 23 | --- | 4 |
| B4 | 2 | 1 | --- | --- | 3 | 3 | 1 | 19 | 2 | 4 |
| B6 | 3 | 1 | --- | --- | 117 | 2 | 1 | --- | --- | 3 |
| B8 | 1 | 2 | --- | --- | --- | 1 | 2 | 59 | --- | 7 |
| B18 | 3 | 2 | 4 | --- | 1 | 2 | 3 | 8 | 1 | 4 |
| B21 | 2 | 1 | --- | --- | --- | 2 | 1 | --- | --- | 4 |
| B22 | 1 | 2 | --- | --- | --- | 1 | 2 | --- | --- | 5 |
| B23 | 2 | 1 | --- | --- | --- | 1 | 2 | --- | --- | 4 |
| B24 | 2 | 1 | --- | --- | --- | 3 | 1 | --- | --- | 2 |
| C1 | 3 | 1 | --- | --- | --- | 2 | 1 | --- | --- | 3 |
| C2 | 2 | 1 | --- | --- | --- | 2 | 1 | --- | --- | 3 |
| C3 | 2 | 1 | --- | --- | --- | 2 | 1 | --- | --- | 3 |
| C4 | 2 | 1 | --- | --- | 79 | 1 | 2 | --- | --- | 3 |
| C5 | 2 | 4 | 3 | 1 | 5 | 2 | 3 | 5 | 1 | 4 |
| C6 | 2 | 1 | --- | --- | --- | 2 | 1 | --- | --- | 3 |
| C7 | 2 | 1 | --- | --- | --- | 2 | 1 | --- | --- | 4 |
| C8 | 2 | 1 | --- | 58 | --- | 1 | 2 | --- | --- | 3 |
| C12 | 3 | 2 | 4 | 1 | --- | 2 | 3 | 5 | 1 | 4 |
| C13 | 3 | 2 | --- | 22 | --- | 2 | 1 | 41 | 4 | 3 |
| C14 | 3 | 2 | --- | 4 | --- | 3 | 1 | 28 | 2 | 4 |
| C15 | 3 | 2 | 88 | 89 | --- | 2 | 1 | 9 | 5 | 3 |
| C18 | 2 | 1 | --- | --- | --- | 2 | 1 | --- | --- | 3 |
| C19 | 3 | 1 | --- | --- | --- | 3 | 1 | 40 | 76 | 2 |
| C21 | 3 | 1 | --- | --- | --- | 3 | 1 | --- | --- | 2 |
| C21 | 2 | 1 | --- | --- | --- | 2 | 1 | --- | --- | 3 |
| C22 | 1 | 2 | 92 | --- | --- | 1 | 2 | --- | --- | 3 |
| C23 | 2 | 1 | --- | --- | --- | 2 | 1 | --- | --- | 3 |

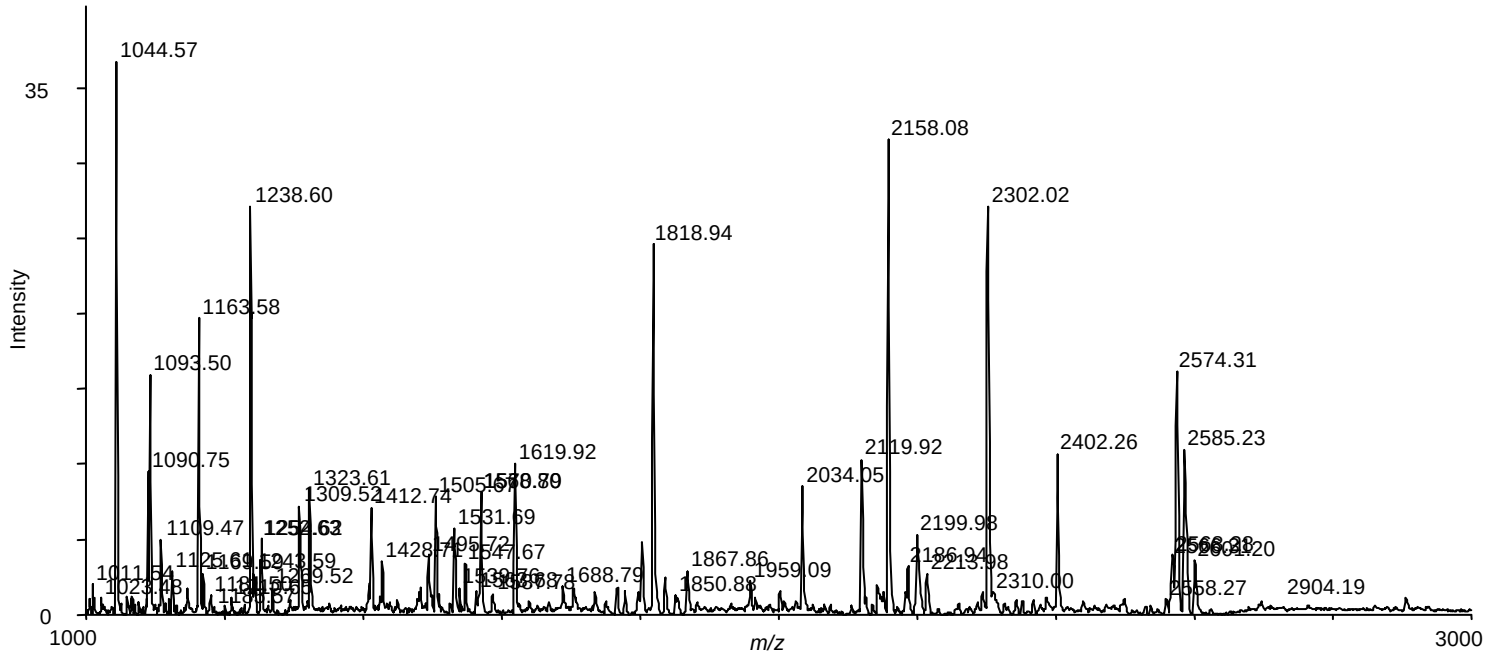
- A random blood sample.
- In 6 cases, BUPID found all 5 chains within the top 20.
- Delta chain are found in all cases within top 10 predictions.
- If the researcher is willing to go down the list, he/she'll be able to find all five chains in BUPID's results.

# Result Comparison: Artificial Spectra (5-Protein Mixture)




- A:  Both MASCOT and BUPID return highly accurate protein identifications.
- B: MASCOT returns accurate result while BUPID gives near-random protein ID.
- C: BUPID returns accurate result while MASCOT gives near or worse than random protein ID.
- ALL SAMPLES are correctly interpreted by BUPID or MASCOT or both.

# E2F1 Protein

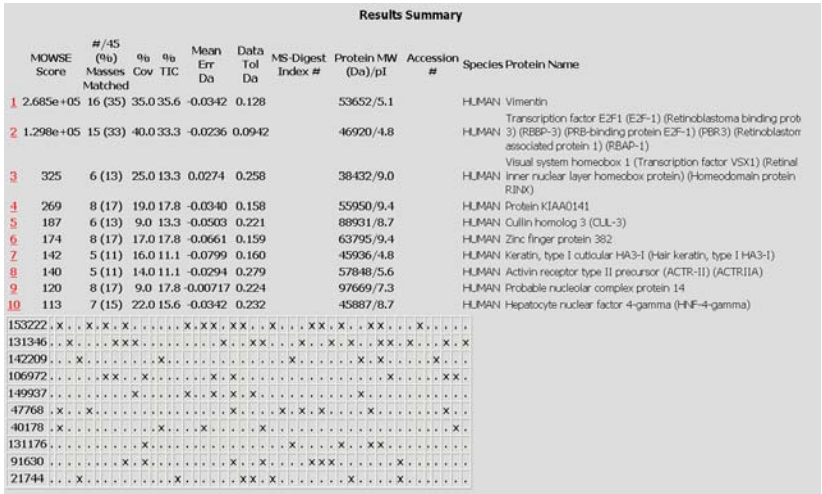



**Search Results** [ show all 7 hits ]

| | Click for Detail | E-value | Raw Score | Matched Peptides | (total | effective) | Sequence Coverage | Protein Sequence Name | | Merge Hits |
|---|---|---|---|---|---|---|---|---|---|---|
| 1. | E2F1_HUMAN | 0.341067 | 99.417922 | 17 | 165 | 94 | 41.7% | Transcription factor E2F1 (E2F-1) (Retinoblasto ... | Q01094 | ☐ |
| 2. | CDN7_HUMAN | 0.403940 | 31.316942 | 5 | 70 | 41 | 25.3% | Cyclin-dependent kinase 4 inhibitor D (P19-INK4D). | P55273 | ☐ |
| 3. | TA6P_HUMAN | 0.409788 | 6.313207 | 1 | 3 | 2 | 18.9% | TAP2-associated 6.5 kDa polypeptide. | Q9Y3F1 | ☐ |
| 4. | OBRG_HUMAN | 0.429088 | 12.373958 | 2 | 30 | 13 | 17.6% | Leptin receptor gene-related protein (OB-R gene ... | O15243 | ☐ |
| 5. | HV3S_HUMAN | 0.430084 | 18.928037 | 3 | 45 | 34 | 35.7% | Ig heavy chain V-III region JON. | P01780 | ☐ |
| 6. | UPA5_HUMAN | 0.430411 | 5.144575 | 1 | 3 | 1 | 100.0% | Unknown protein from 2D-page of plasma (Spot 13 ... | P30091 | ☐ |
| 7. | VIME_HUMAN | 0.432666 | 102.438835 | 16 | 300 | 234 | 37.0% | Vimentin. | P08670 | ☐ |
| | | | | | | | | | Trypsin | ☑ |

Plot

# E2F1 Protein



**Results Summary**

| | MOWSE Score | #/45 (%) Masses Matched | % Cov | % TIC | Mean Err Da | Data Tol Da | MS-Digest Index # | Protein MW (Da)/pI | Accession # | Species | Protein Name |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 2.685e+05 | 16 (35) | 35.0 | 35.6 | -0.0342 | 0.128 | | 53652/5.1 | | HUMAN | Vimentin |
| 2 | 1.298e+05 | 15 (33) | 40.0 | 33.3 | -0.0236 | 0.0942 | | 46920/4.8 | | HUMAN | Transcription factor E2F1 (E2F-1) (Retinoblastoma binding protein 3) (RBBP-3) (PRB-binding protein E2F-1) (PBR3) (Retinoblastoma-associated protein 1) (RBAP-1) |
| 3 | 325 | 6 (13) | 25.0 | 13.3 | 0.0274 | 0.258 | | 38432/9.0 | | HUMAN | Visual system homeobox 1 (Transcription factor VSX1) (Retinal inner nuclear layer homeobox protein) (Homeodomain protein RINX) |
| 4 | 269 | 8 (17) | 19.0 | 17.8 | -0.0340 | 0.158 | | 55950/9.4 | | HUMAN | Protein KIAA0141 |
| 5 | 187 | 6 (13) | 9.0 | 13.3 | -0.0503 | 0.221 | | 88931/8.7 | | HUMAN | Cullin homolog 3 (CUL-3) |
| 6 | 174 | 8 (17) | 17.0 | 17.8 | -0.0661 | 0.159 | | 63795/9.4 | | HUMAN | Zinc finger protein 382 |
| 7 | 142 | 5 (11) | 16.0 | 11.1 | -0.0799 | 0.160 | | 45936/4.8 | | HUMAN | Keratin, type I cuticular HA3-I (Hair keratin, type I HA3-I) |
| 8 | 140 | 5 (11) | 14.0 | 11.1 | -0.0294 | 0.279 | | 57848/5.6 | | HUMAN | Activin receptor type II precursor (ACTR-II) (ACTRIIA) |
| 9 | 120 | 8 (17) | 9.0 | 17.8 | -0.00717 | 0.224 | | 97669/7.3 | | HUMAN | Probable nucleolar complex protein 14 |
| 10 | 113 | 7 (15) | 22.0 | 15.6 | -0.0342 | 0.232 | | 45887/8.7 | | HUMAN | Hepatocyte nuclear factor 4-gamma (HNF-4-gamma) |

153222 . x . . x . x . . . . . x . x x . x x . . x . . . x x . x . . x x . . . . .
131346 . . x . . . . x x x . . . . . . . x . . x x . . . . x . . x . x . . x x . x . . . x . x
142209 . . . x . . . . . . . x . . . . . . . . . . . x . . . . . x . . x . . . . x . . .
106972 . . . . . x x . . . x . . . . . . x . x . . . . . . . . . . . . x . . . . . x x .
149937 . . . . . . . . x . . . . . x . . x . . x . x . . . . . . . . . x . . . . . . . .
47768 . x . . x . . . . . . . . . . . . . . x . . . . x . x . x . . . . x . . . . . . x . .
40178 . x . . . . . . . . . . x . . . . x . . . . . x . . . . . . . . . . . . . . . . x .
131176 . . . . . . . . . . x . . . . . . . . . . . . . . x . . . . x . . x x . . . . . . .
91630 . . . . . . . x . x . . . . . . . . x . . x . . . . x x x . . . . . . x . . . . . .
21744 . . . x . . . . . . . . x . . . . . x x . x . . . . . . x . . . . x . . . . . . .

MS-FIT

**Concise Protein Summary Report**

Format As: Concise Protein Summary — Help

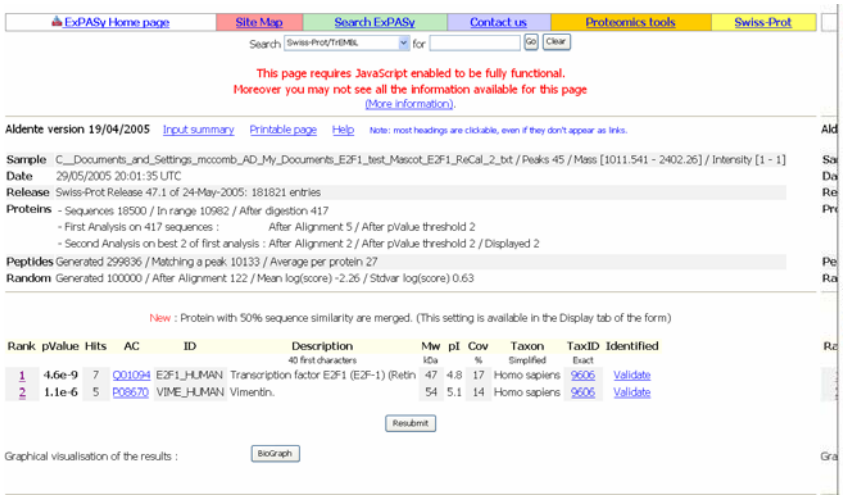
Significance threshold p< 0.05 — Max. number of hits 20

Re-Search All — Search Unmatched

1. **Mixture 1** **Total score: 217** **Expect:** 1e-17 **Queries matched:** 29
Components (only one family member shown for each component):
Q01094-00-00-00 **Mass:** 46891 **Score:** 123 **Expect:** 2.6e-08 **Queries matched:** 15
(E2F1_HUMAN) Splice isoform Displayed; Variant Displayed; Conflict Displayed; from Q01094 Transcription factor E2F1 (E2F-1) (Retinoblastoma binding
P08670 **Mass:** 53488 **Score:** 98 **Expect:** 8.8e-06 **Queries matched:** 16
(VIME_HUMAN) Vimentin Vimentin

2. Q01094-00-00-00 **Mass:** 46891 **Score:** 123 **Expect:** 2.6e-08 **Queries matched:** 15
(E2F1_HUMAN) Splice isoform Displayed; Variant Displayed; Conflict Displayed; from Q01094 Transcription factor E2F1 (E2F-1) (Retinoblastoma binding
Q01094-00-04-00 **Mass:** 46921 **Score:** 123 **Expect:** 2.6e-08 **Queries matched:** 15
(E2F1_HUMAN) Splice isoform Displayed; Variant dbSNP:3213176; Conflict Displayed; from Q01094 Transcription factor E2F1 (E2F-1) (Retinoblastoma bind
Q01094-00-03-00 **Mass:** 46904 **Score:** 111 **Expect:** 4e-07 **Queries matched:** 14
(E2F1_HUMAN) Splice isoform Displayed; Variant dbSNP:3213174; Conflict Displayed; from Q01094 Transcription factor E2F1 (E2F-1) (Retinoblastoma bind
Q01094-00-02-00 **Mass:** 46923 **Score:** 110 **Expect:** 5.1e-07 **Queries matched:** 14
(E2F1_HUMAN) Splice isoform Displayed; Variant dbSNP:3213173; Conflict Displayed; from Q01094 Transcription factor E2F1 (E2F-1) (Retinoblastoma bind
Q01094-00-01-00 **Mass:** 46872 **Score:** 99 **Expect:** 6.7e-06 **Queries matched:** 13
(E2F1_HUMAN) Splice isoform Displayed; Variant dbSNP:3213172; Conflict Displayed; from Q01094 Transcription factor E2F1 (E2F-1) (Retinoblastoma bind

MASCOT



ExPASy Home page | Site Map | Search ExPASy | Contact us | Proteomics tools | Swiss-Prot

Search Swiss-Prot/TrEMBL for — Go — Clear

This page requires JavaScript enabled to be fully functional.
Moreover you may not see all the information available for this page
(More information).

**Aldente version 19/04/2005** Input summary | Printable page | Help Note: most headings are clickable, even if they don't appear as links.

**Sample** C__Documents_and_Settings_mccomb_AD_My_Documents_E2F1_test_Mascot_E2F1_ReCal_2_txt / Peaks 45 / Mass [1011.541 - 2402.26] / Intensity [1 - 1]
**Date** 29/05/2005 20:01:35 UTC
**Release** Swiss-Prot Release 47.1 of 24-May-2005: 181821 entries
**Proteins** - Sequences 18500 / In range 10982 / After digestion 417
- First Analysis on 417 sequences : After Alignment 5 / After pValue threshold 2
- Second Analysis on best 2 of first analysis : After Alignment 2 / After pValue threshold 2 / Displayed 2
**Peptides** Generated 299836 / Matching a peak 10133 / Average per protein 27
**Random** Generated 100000 / After Alignment 122 / Mean log(score) -2.26 / Stdvar log(score) 0.63

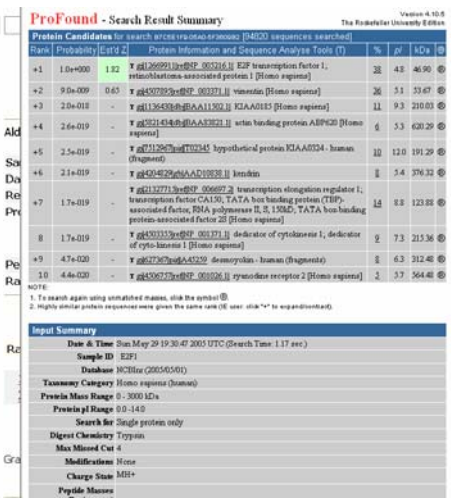
New : Protein with 50% sequence similarity are merged. (This setting is available in the Display tab of the form)

| Rank | pValue | Hits | AC | ID | Description (40 first characters) | Mw kDa | pI | Cov % | Taxon Simplified | TaxID Exact | Identified |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 4.6e-9 | 7 | Q01094 | E2F1_HUMAN | Transcription factor E2F1 (E2F-1) (Retin | 47 | 4.8 | 17 | Homo sapiens | 9606 | Validate |
| 2 | 1.1e-6 | 5 | P08670 | VIME_HUMAN | Vimentin. | 54 | 5.1 | 14 | Homo sapiens | 9606 | Validate |

Resubmit

Graphical visualisation of the results : BioGraph

Aldante



**ProFound** - Search Result Summary

**Protein Candidates** for search [redacted] [04020 sequences searched]

| Rank | Probability | Est'd Z | Protein Information and Sequence Analyse Tools (T) | % | pI | kDa |
|---|---|---|---|---|---|---|
| +1 | 1.0e+000 | 1.82 | gi\|12669911\|ref\|NP_003216.1\| E2F transcription factor 1; retinoblastoma-associated protein 1 [Homo sapiens] | 38 | 4.8 | 46.90 |
| +2 | 9.0e-009 | 0.65 | gi\|4507289\|ref\|NP_003371.1\| vimentin [Homo sapiens] | 36 | 5.1 | 53.67 |
| +3 | 2.0e-018 | - | gi\|11364308\|dbj\|BAA11502.1\| KIAA0183 [Homo sapiens] | 11 | 9.3 | 210.03 |
| +4 | 2.6e-019 | - | gi\|5821434\|dbj\|BAA83821.1\| actin binding protein ABP620 [Homo sapiens] | 6 | 5.3 | 620.29 |
| +5 | 2.5e-019 | - | gi\|7112967\|pir\|T02345 hypothetical protein KIAA0324 - human (fragment) | 10 | 12.0 | 191.29 |
| +6 | 2.1e-019 | - | gi\|4240429\|gb\|AAD10838.1\| keratin | 8 | 5.4 | 376.32 |
| +7 | 1.7e-019 | - | gi\|2132771\|pir\|\|JC6097 transcription elongation regulator 1; transcription factor CA150; TATA box binding protein (TBP)-associated factor, RNA polymerase II, 150kD; TATA box binding protein-associated factor 2B [Homo sapiens] | 14 | 8.8 | 123.88 |
| 8 | 1.7e-019 | - | gi\|4503335\|ref\|NP_001371.1\| dedicator of cytokinesis 1; dedicator of cyto-kinesis 1 [Homo sapiens] | 9 | 7.3 | 215.36 |
| +9 | 4.7e-020 | - | gi\|627367\|gb\|AA52259 desmoyokin - human (fragment) | 8 | 6.3 | 312.48 |
| 10 | 4.4e-020 | - | gi\|4506757\|ref\|NP_003226.1\| synaptotagmin receptor 2 [Homo sapiens] | 3 | 3.7 | 564.48 |

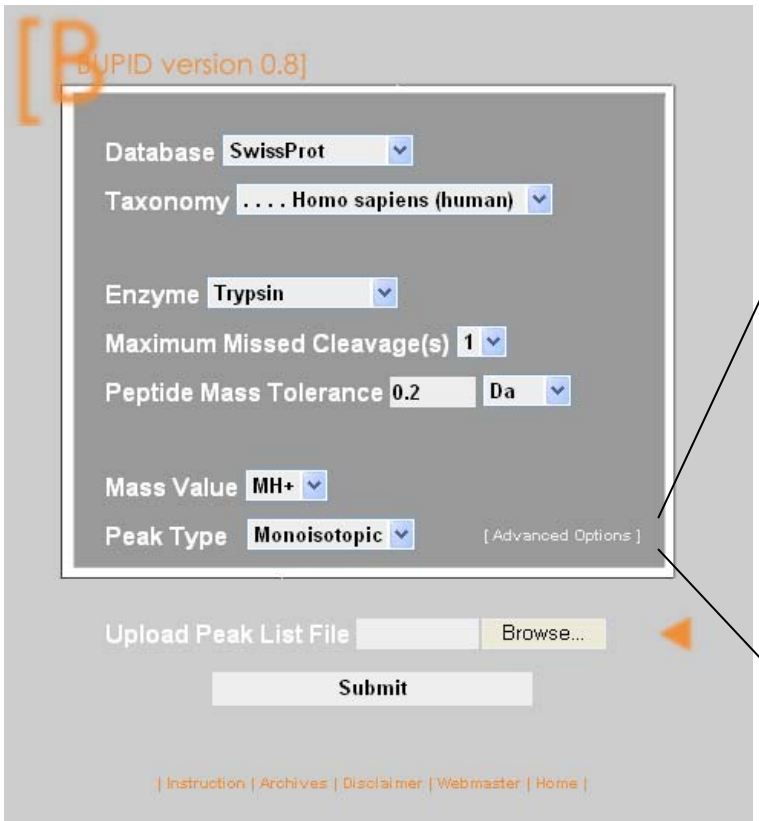
NOTE:
1. To search again using unmatched masses, click the symbol Ⓜ.
2. Highly similar protein sequences were given the same rank (IE user: click "+" to expand/contract).

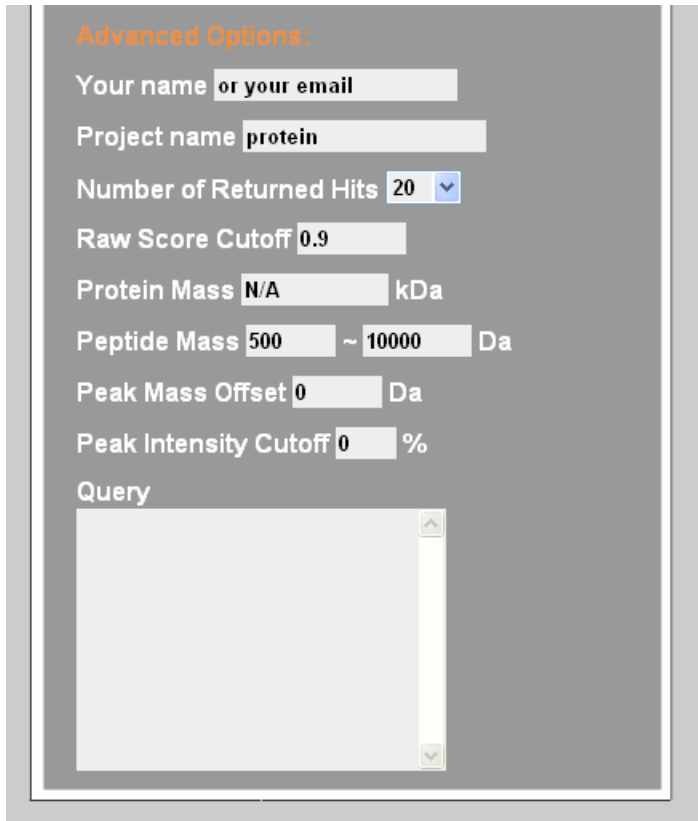
**Input Summary**

| | |
|---|---|
| **Date & Time** | Sun May 29 19:30:47 2005 UTC (Search Time: 1.17 sec) |
| **Sample ID** | E2F1 |
| **Database** | NCBInr (2005/05/01) |
| **Taxonomy Category** | Homo sapiens (human) |
| **Protein Mass Range** | 0 - 3000 kDa |
| **Protein pI Range** | 0.0 -14.0 |
| **Search for** | Single protein only |
| **Digest Chemistry** | Trypsin |
| **Max Missed Cut** | 4 |
| **Modifications** | None |
| **Charge State** | MH+ |
| **Peptide Masses** | |

ProFound

# BUPID Web Interface




The standard interface of BUPID offers standard parameters.

Users also have access to more options one mouse-click away.

[Advanced Options]

## Search Results

[ show 5 more hits ] [ show all 169 hits ] [ help ]

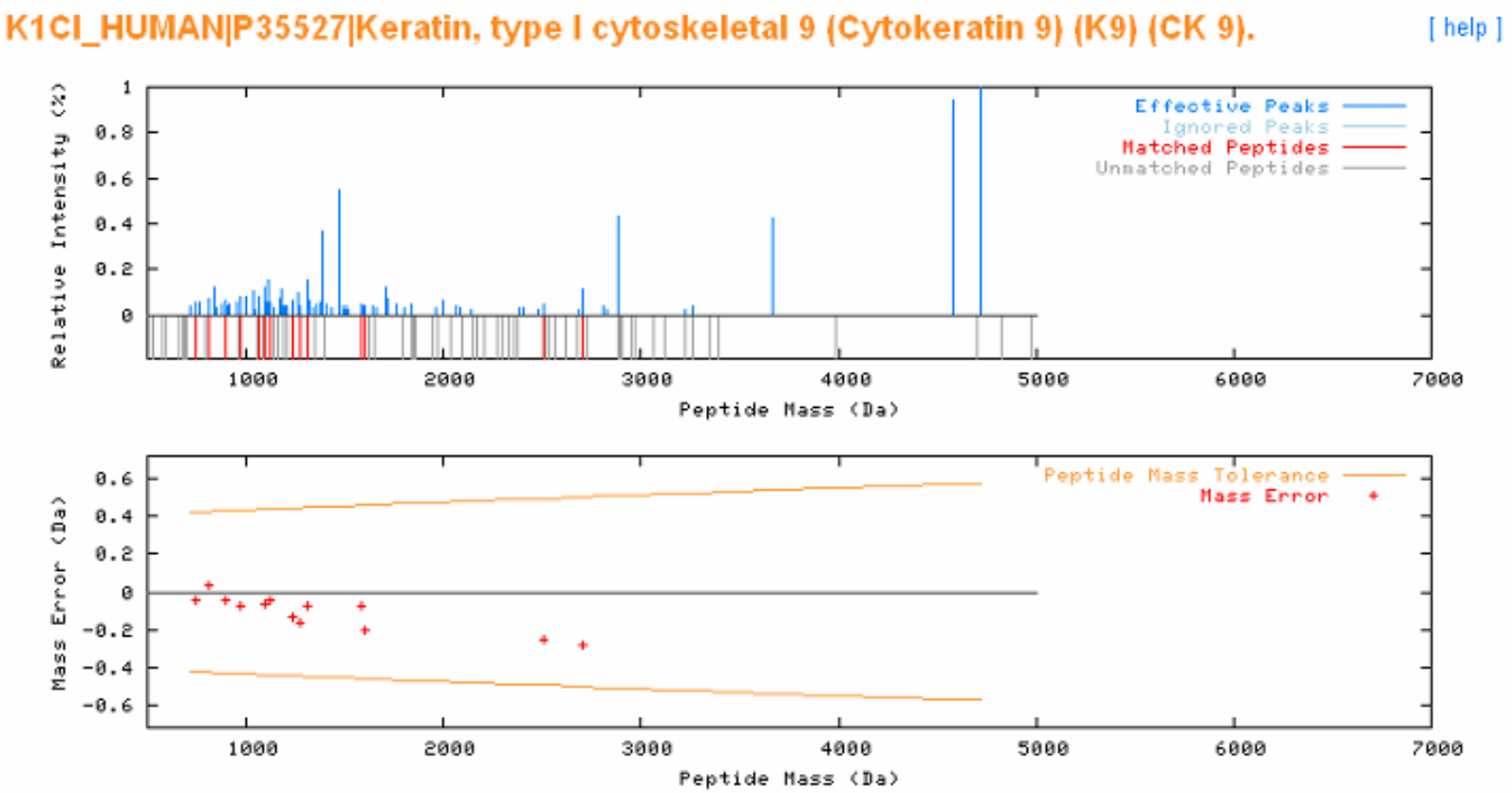
| | Click for Detail | score | Raw Score | Matched Peptides | (total | effective) | Sequence Coverage | Protein Sequence Name | | Merge Hits |
|---|---|---|---|---|---|---|---|---|---|---|
| 1. | RS30_HUMAN | 0.275882 | 16.704674 | 3 | 39 | 16 | 39.6% | 40S ribosomal protein S30. | Q05472 | ☐ |
| 2. | K1CI_HUMAN | 0.333126 | 75.270339 | 14 | 99 | 79 | 24.9% | Keratin, type I cytoskeletal 9 (Cytokeratin 9) ... | P35527 | ☐ |
| 3. | K1CJ_HUMAN | 0.343009 | 65.151715 | 12 | 91 | 72 | 22.3% | Keratin, type I cytoskeletal 10 (Cytokeratin 10 ... | P13645 | ☐ |
| 4. | K2C1_HUMAN | 0.344604 | 88.926094 | 18 | 129 | 100 | 33.4% | Keratin, type II cytoskeletal 1 (Cytokeratin 1) ... | P04264 | ☐ |
| 5. | CAHA_HUMAN | 0.351723 | 38.177582 | 7 | 53 | 43 | 15.9% | Carbonic anhydrase-related protein 10 (Carbonic ... | Q9NS85 | ☐ |
| 6. | H13_HUMAN | 0.363524 | 27.715259 | 6 | 113 | 57 | 18.6% | Histone H1.3 (Histone H1c). | P16402 | ☐ |
| 7. | RL31_HUMAN | 0.367282 | 21.808564 | 4 | 61 | 36 | 32.8% | 60S ribosomal protein L31. | P12947 | ☐ |
| 8. | Z125_HUMAN | 0.367869 | 21.903299 | 4 | 33 | 24 | 25.0% | Zinc finger protein 125 (HZF-3) (Fragment). | P35274 | ☐ |
| 9. | RL44_HUMAN | 0.370291 | 16.519316 | 3 | 63 | 28 | 31.3% | 60S ribosomal protein L44 (L36a). | P09896 | ☐ |
| 10. | HV1C_HUMAN | 0.375541 | 17.414945 | 3 | 21 | 19 | 34.2% | Ig heavy chain V-I region ND precursor (Fragmen ... | P01744 | ☐ |
| 11. | P628_HUMAN | 0.379520 | 11.876338 | 2 | 15 | 12 | 71.1% | Protein PRO0628. | Q9UI54 | ☐ |
| 12. | BAXC_HUMAN | 0.381664 | 10.699630 | 2 | 11 | 10 | 57.5% | BAX protein, cytoplasmic isoform gamma. | Q07815 | ☐ |
| 13. | BUB3_HUMAN | 0.383651 | 39.692537 | 7 | 67 | 51 | 21.6% | Mitotic checkpoint protein BUB3. | O43684 | ☐ |
| 14. | RS17_HUMAN | 0.384568 | 22.382246 | 5 | 53 | 31 | 22.4% | 40S ribosomal protein S17. | P08708 | ☐ |
| 15. | S113_HUMAN | 0.386341 | 17.076007 | 3 | 33 | 20 | 26.8% | S100 calcium-binding protein A13. | Q99584 | ☐ |
| 16. | SR19_HUMAN | 0.387692 | 22.615362 | 4 | 55 | 33 | 29.2% | Signal recognition particle 19 kDa protein (SRP ... | P09132 | ☐ |
| 17. | ARGR_HUMAN | 0.389733 | 32.298601 | 6 | 107 | 55 | 26.1% | Arginine-rich protein. | P55145 | ☐ |
| 18. | K1CX_HUMAN | 0.390709 | 49.354253 | 11 | 93 | 75 | 16.0% | Keratin, type I cytoskeletal 17 (Cytokeratin 17 ... | P08779 | ☐ |
| 19. | IG2R_HUMAN | 0.391973 | 11.910041 | 2 | 15 | 12 | 51.6% | Putative insulin-like growth factor II associat ... | P09565 | ☐ |
| 20. | TRIF_HUMAN | 0.393250 | 26.328247 | 5 | 79 | 55 | 19.9% | Troponin I, fast skeletal muscle (Troponin I, f ... | P48788 | ☐ |
| | | | | | | | | | Trypsin | ☑ |

Plot

Final results are ranked by the **score**. Click on the link in the first column to view each individual proteins. Check boxes in the last column and click **Plot** for more data-processing (see bellow).

Protein view provides

- A plot of spectrum and matched theoretical peaks
- Sequence and coverage
- Matched and miss matched peptides



## Sequence

```
  1 MSCRQFSSSY LTSGGGGGGG LGSGGSIRSS YSRFSSSGGR GGGGRFSSSS
 51 GYGGGSSRVC GRGGGGSFGY SYGGGSGGGF SASSLGGGFG GGSRGFGGAS
101 GGGYSSSGGF GGGFGGGSGG GFGGGYGSGF GGLGGFGGGA GGGDGGILTA
151 NEKSTMQELN SRLASYLDKV QALEEANNDL ENKIQDWYDK KGPAAIQKNY
201 SPYYNTIDDL KDQIVDLTVG NNKTLLDIDN TRMTLDDFRI KFEMEQNLRQ
```

… … etc.

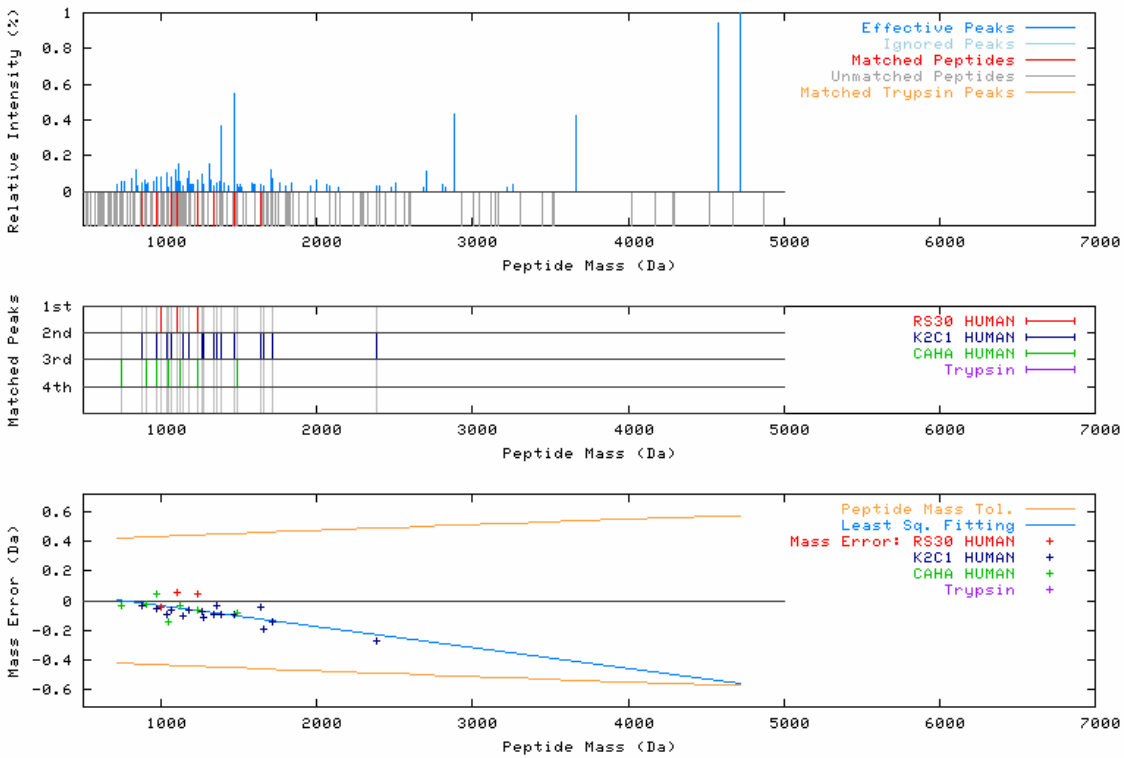
## Matched Peptides

| Start - | End | Length | Peptide Mass | Peak Mass | Mass Error | Missed Cl. | Peptide Sequence |
|---|---|---|---|---|---|---|---|
| 1 - | 4 | 4 | 495.193379 | | | 0 | MSCR |
| 1 - | 28 | 28 | 2622.191483 | | | 1 | MSCRQFSSSYLTSGGGGGGGLGSGGSIR |
| 5 - | 28 | 24 | 2145.008674 | | | 0 | QFSSSYLTSGGGGGGGLGSGGSIR |
| 5 - | 33 | 29 | 2725.269201 | | | 1 | QFSSSYLTSGGGGGGGLGSGGSIRSSYSR |
| 29 - | 33 | 5 | 598.271097 | | | 0 | SSYSR |
| 29 - | 40 | 12 | 1276.579632 | 1276.741975 | -0.162343 | 1 | SSYSRFSSSGGR |
| 34 - | 40 | 7 | 696.319105 | | | 0 | FSSSGGR |
| 34 - | 45 | 12 | 1080.506071 | | | 1 | FSSSGGRGGGGR |
| 41 - | 45 | 5 | 402.197536 | | | 0 | GGGGR |
| 41 - | 58 | 18 | 1618.708416 | | | 1 | GGGGRFSSSSGYGGGSSR |
| 46 - | 58 | 13 | 1234.521450 | 1234.659975 | -0.138525 | 0 | FSSSSGYGGGSSR |
| 46 - | 62 | 17 | 1649.721628 | | | 1 | FSSSSGYGGGSSRVCGR |
| 59 - | 62 | 4 | 433.210748 | | | 0 | VCGR |
| 59 - | 94 | 36 | 3119.354019 | | | 1 | VCGRGGGGSFGYSYGGGSGGGFSASSLGGGFGGGSR |
| 63 - | 94 | 32 | 2704.153841 | 2704.437975 | -0.284134 | 0 | GGGGSFGYSYGGGSGGGFSASSLGGGFGGGSR |
| 63 - | 153 | 91 | 7513.231551 | | | 1 | GGGGSFGYSYGGGSGGGFSASSLGGGFGGGSRGFGGASGGG |
| 95 - | 153 | 59 | 4827.088280 | | | 0 | GFGGASGGGYSSSGGFGGGFGGGSGGGFGGGYGSGFGGLGG |

… … etc.

# Protein Mixture View

**Spectrum coverage: 35.4%**




**Protein Mixture**

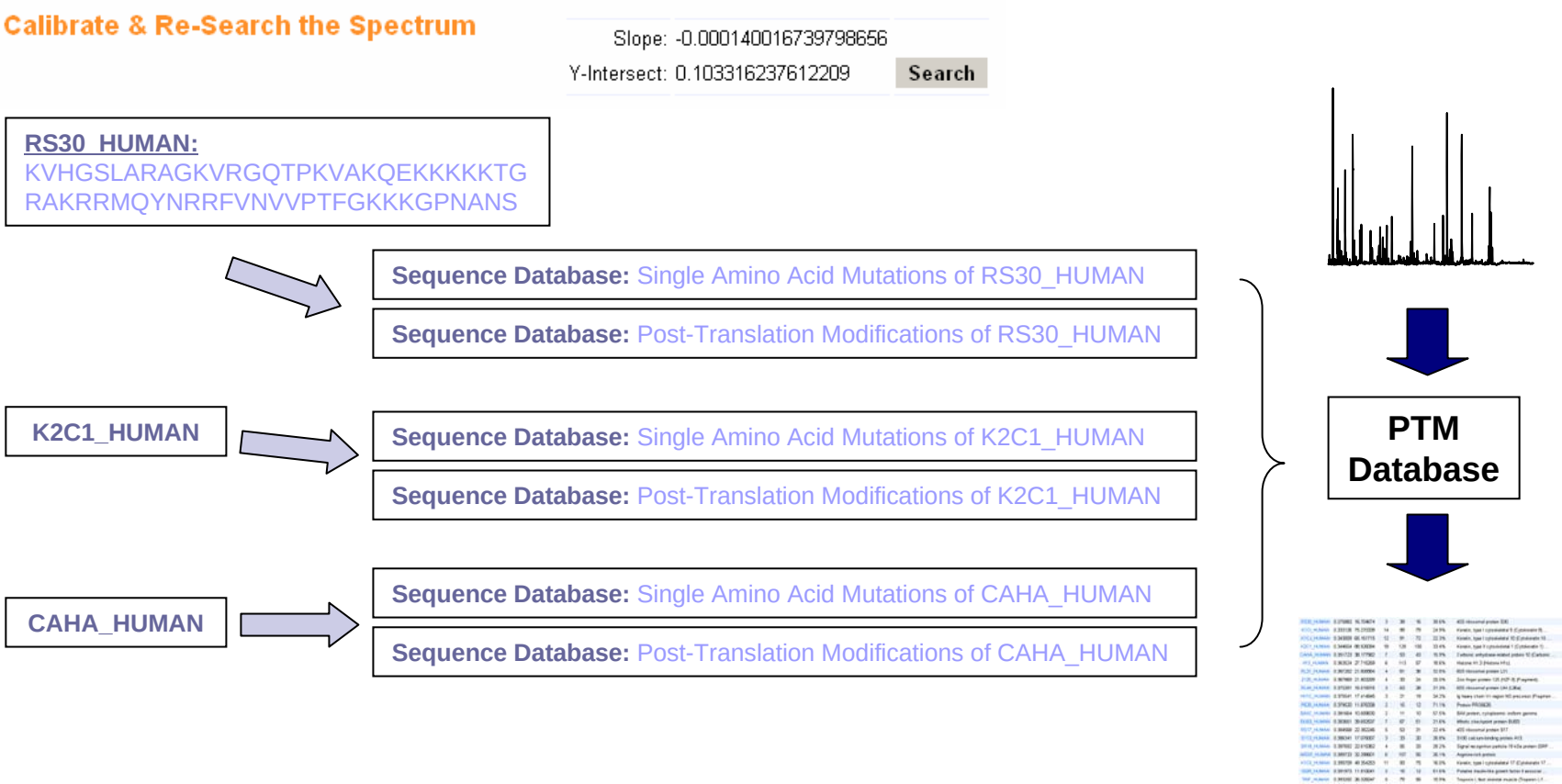
| | E-value | Raw Score | Matched Peptides | (total | effective) | Sequence Coverage | Protein Sequence Name | |
|---|---|---|---|---|---|---|---|---|
| RS30_HUMAN | 0.275882 | 16.704674 | 3 | 39 | 16 | 39.6% | 40S ribosomal protein S30. | Q05472 |
| K2C1_HUMAN | 0.344604 | 88.926094 | 18 | 129 | 100 | 33.4% | Keratin, type II cytoskeletal 1 (Cytokeratin 1) ... | P04264 |
| CAHA_HUMAN | 0.351723 | 38.177582 | 7 | 53 | 43 | 15.9% | Carbonic anhydrase-related protein 10 (Carbonic ... | Q9NS85 |
| | | | | | | | | Trypsin |

- Protein mixture view plots all matched and miss matched theoretical peaks along with the spectrum. Users also have the choice to show peaks of self-digestion (peptides of the restriction enzyme).

- All matched peaks are also plotted back to back for users to identify shared peptides or peptides with similar masses. In popular search engines such as MASCOT, matched peaks are removed after each round. Thus peptides overlap with another peptide with a stronger signal may not be picked up in the database search. BUPID doesn't remove peaks and is free from such occasions.

- Mass errors of all peptides are plotted together. A trend line generated from least-square-fit helps to identify the trend of the error. If, for instance, that the researcher believes that the trend line resembles the system error of the machine, BUPID offers an internal calibration option to adjust for such error.

# Search for Post-Translation Modifications




- BUPID can search for post-translation modifications within a set of user specified proteins.
- For each protein sequence, BUPID creates a database with post-translation modification and variations, against which the spectrum is searched again. Final results are ranked by the p-value of their log-likelihood score.

## Acknowledgement

This project has been funded in whole or in part with Federal funds as part of the NHLBI Proteomics Initiative from the National Heart, Lung, and Blood Institute, National Institutes of Health, under Contract No. N01-HV-28178. We are also grateful for technological support from Bruker Daltonics.