4 Ergebnisse und Diskussion

Die unten beschriebenen Algorithmen und das von mir entwickelte Softwarepaket MS-Proteomics wurden anhand umfangreicher Datensätze, die mit der zuvor beschriebenen Hochdurchsatztechnik (siehe 1.4) generiert wurden, erprobt. Insbesondere ca. 1000 verschiedene, gentechnisch hergestellte rekombinante Proteine aus der in der Abteilung Lehrach am MPI für molekulare Genetik generierten cDNA Expressionsbibliothek von fötalem menschlichem Gehirn [75] und ca. 40 rekombinierte Proteine von „*Arabidopsis Thaliana*“, deren Identität und Größe unabhängig mittels DNA-Sequenzierung eindeutig bestimmt wurde [76], wurden mit Trypsin gespalten und mehrfach analysiert. Diese Daten wurden zur Entwicklung und Verifizierung der Algorithmen verwendet.

4.1 Algorithmen zur Verbesserung der Massenrichtigkeit

Die von mir entwickelten und nachfolgend beschriebenen Algorithmen basieren auf der Beobachtung, dass eine Verschlechterung der Massenrichtigkeit im wesentlichen auf zwei systematische Fehler zurückzuführen ist. Bei der Nullpunktverschiebung des Massenspektrums weichen alle Massen um einen bestimmten Betrag (im Extremfall bis zu mehreren Dalton) von ihren erwarteten Werten ab. Dieser Effekt behaftet also jede einzelne gemessene Masse mit einem bestimmten für alle Massen gleichen Fehlerbetrag. Die Differenzen zwischen den einzelnen gemessenen Massen bleiben davon jedoch unberührt [55].
Bei der Streckung oder Stauchung des Massenspektrums weichen die gemessenen Massen um unterschiedliche Beträge von ihren erwarteten Werten ab. Die Beträge der Abweichung nehmen bei größeren Massen zu. Bei diesem Effekt werden sowohl die Absolutwerte, als auch die Massendifferenzen verändert.

Im Rahmen meiner Diplomarbeit [54] konnte ich zeigen, dass beim Wechsel der Probenposition bei sonst konstanten experimentellen Bedingungen die hieraus resultierende Streckung oder Stauchung der Spektren entlang der m/z-Achse linear ist, d.h. sie kann durch eine einfache Transformations-gleichung der Form: $\mathbf{m_{korrekt} = a \bullet m_{gemessen} + b}$ korrigiert werden.

Hiermit habe ich folgenden Lösungsansatz entwickelt:

1. In einem ersten Schritt werden für das erste in der ausgewählten Sequenzdatenbank eingetragene Protein die Massen für alle theoretisch möglichen Peptide, die beim Verdau dieses Proteins durch das verwendete Enzym (z.B. Trypsin) entstehen können, berechnet. Alle gemessenen Massen, die in einem großen Fehlerintervall (z.B. $\pm$ 500ppm), mit den theoretisch berechneten Massen übereinstimmen werden im weiteren berücksichtigt, alle anderen verworfen.
Die zugelassene Fehlertoleranz gewährleistet, dass alle möglichen Kandidaten erfasst werden und ist nötig, da bei Verwendung einer externen Kalibrierung eine Nullpunktverschiebung bis zu 0,5 Da möglich ist.

2. Als zweiter Schritt wird die Abweichung jeder experimentell bestimmten Masse von ihrer korrespondierenden, erwarteten Peptidmasse berechnet und anschließend der Mittelwert **μ** und die Standard-abweichung σ für alle Abweichungen ermittelt. In den nachfolgenden Berechnungen werden nur die Massen berücksichtigt für deren Abweichung gilt:

$$\mathbf{m \leq \mu \pm 2\sigma.} \qquad \mathbf{(5)}$$

Dadurch werden Ausreißer, die z.B. durch fehlerhaftes „Peakpicking“ (z.B. kann aufgrund eines schlechten Signal-zu-Rauschverhältnisses

im Spektrum der erste Isotopenpeak im Rauschen verloren gehen, so dass stattdessen fälschlicherweise der zweite Isotopenpeak gepickt wird) vorhanden sein können, von den weiteren Berechnungen ausgeschlossen.

3. Im dritten Schritt werden die Abweichungen der experimentell bestimmten Massen gegen die korrespondierenden, berechneten Peptidmassen geplottet und die Regressionskoeffizienten der einfachen linearen Regression berechnet, mit deren Hilfe die im zweiten Schritt für jede experimentell bestimmte Masse berechneten Abweichungen korrigiert werden (siehe unten).

4. In einem vierten Schritt wird der Mittelwert **μ** und die Standardabweichung σ für alle nunmehr korrigierten Abweichungen ermittelt und ggf. weitere Ausreißer eliminiert, indem im folgenden nur Massen berücksichtigt werden, deren Abweichungen Gleichung 5 gehorchen. Alle anderen Massen bilden in ihrer Summe die Trefferanzahl (Hits) für dieses Protein. Aus deren Abweichungen wird die endgültige Standardabweichung σ ermittelt.

5. Im letzten Schritt werden für alle unter 4 ermittelten Treffer die Längen der Aminosäuresequenzen der jeweiligen Peptide addiert (identische Aminosäuresequenzen werden nur einmal berücksichtigt) und aus der resultierenden Summe $L_{Peptide}$ und der Länge der Aminosäuresequenz des gesamten Proteins $L_{Protein}$ die prozentuale Sequenzabdeckung des Proteins **SC** wie folgt ermittelt:

$$SC = \frac{L_{Peptide}}{L_{Protein}} \times 100 \qquad \textbf{(6)}$$

Schritte 1-5 werden anschließend für alle weiteren in der ausgewählten Sequenzdatenbank eingetragenen Proteine wiederholt.

Im Folgenden wird die Effizienz der Algorithmen an einem konkreten Beispiel veranschaulicht. Eine Probe der tryptisch verdauten Peptide des gentechnisch hergestellten menschlichen Proteins „*Actin, Cytoplasmic 1 (Beta-Actin)*“ wurde an der Position G16, der für die Kalibrierung verwendete Peptidstandard I (siehe 3.1,Tabelle 4 und 3.2.3) in unmittelbarer Nähe an der Position G12 des verwendeten 384er MALDI-Probenträgers aufgetragen (Abbildung 14).

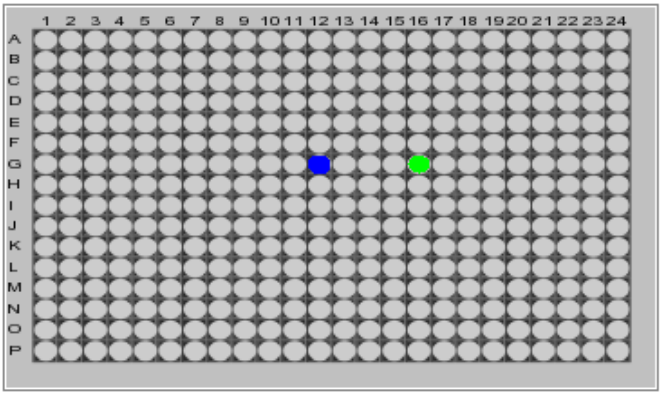

Abbildung 14 Schematische Darstellung des 384er MALDI-Probenträgers.
Der Peptidstandard wurde an Position G12 (blau markiert) aufgetragen, die Probe an Position G16 (grün markiert).

Die Probenpräparation wurde wie unter 3.2.2, das „*Peakpicking*“ und die Massenbestimmung wie unter 3.2.1 beschrieben durchgeführt. Die resultierende Peakliste enthielt 37 Einträge, die entsprechenden Einträge sind im Massenspektrum des Proteins (Abbildung 15) grün bzw. rot eingefärbt sind. Grün gefärbte Peaks entsprechen tryptischen Peptiden des gesuchten Proteins, rot gefärbte nicht.

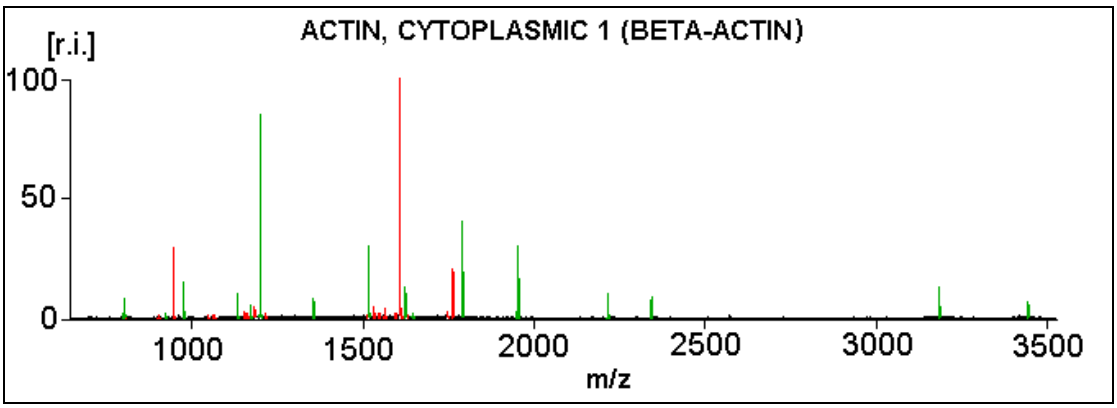

Abbildung 15 MALDI-TOF-Massenspektrum des tryptisch verdauten rekombinierten menschlichen Protein „*Actin, Cytoplasmic 1 (Beta-Actin)*“.
Grün eingefärbte Peaks gehören zu „*Actin, Cytoplasmic 1 (Beta-Actin)* “, rot eingefärbte Peaks nicht.

In Abbildung 16 sind die Abweichungen in ppm aller experimentell bestimmten Massen von den für das untersuchte Protein berechneten Peptidmassen, die in einem Fehlerintervall von ± 500 ppm mit diesen übereinstimmen, versus den theoretisch berechneten Peptidmassen aufgetragen. Analog zu Schritt 1 (siehe oben) wurden dem untersuchten Protein 21 von 37 möglichen Peptidmassen zugeordnet.

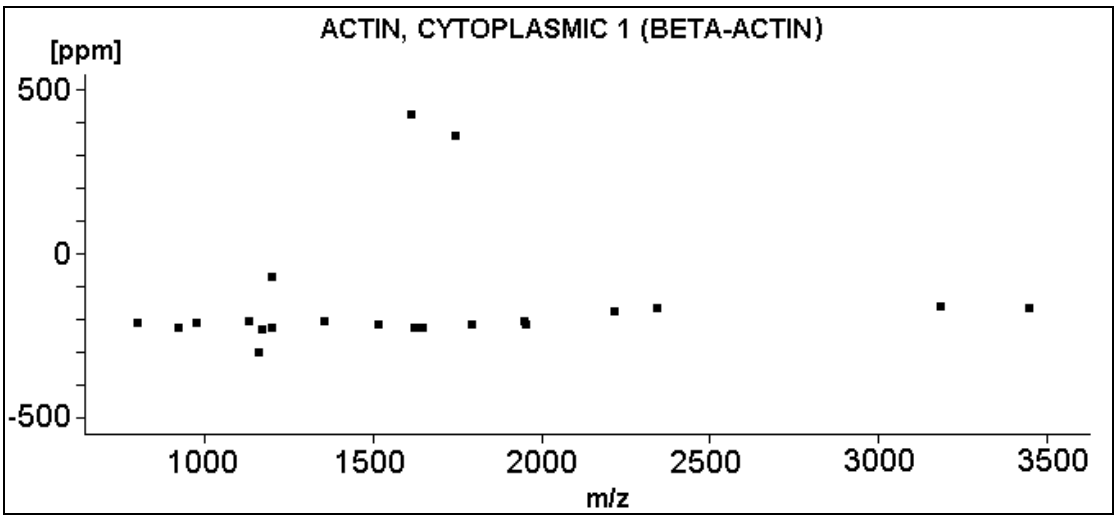

Abbildung 16 Realtive Abweichungen in ppm der experimentell bestimmten Massen von den für das Protein „*Actin, Cytoplasmic 1 (Beta-Actin)*“ berechneten Peptidmassen, die in in einem Fehlerintervall von ± 500 ppm mit diesen übereinstimmen, versus den berechneten Peptidmassen.

Die ermittelten Abweichungen liegen in einem Bereich von –299 ppm bis +430 ppm. Der berechnete Mittelwert **μ** beträgt –144,9 ppm und die ermittelte Standardabweichung σ ist 180,7ppm. Daraus ergibt sich analog zu Schritt 2 für die untere Grenze ein erlaubter Fehler **μ-2σ** von –506,3 ppm und für die obere Grenze ein erlaubter Fehler **μ+2σ** von +216,6 ppm. Die beiden roteingefäbten und mit Pfeil markierten Peptidmassen in Abbildung 17 besitzen eine Abweichung die außerhalb dieser erlaubten Grenzen liegt (+430 ppm bzw. 363 ppm) und werden deshalb von der weiteren Berechnung ausgeschlossen. Dem untersuchten Protein wurden nunmehr noch 19 von 37 möglichen Peptidmassen zugeordnet.

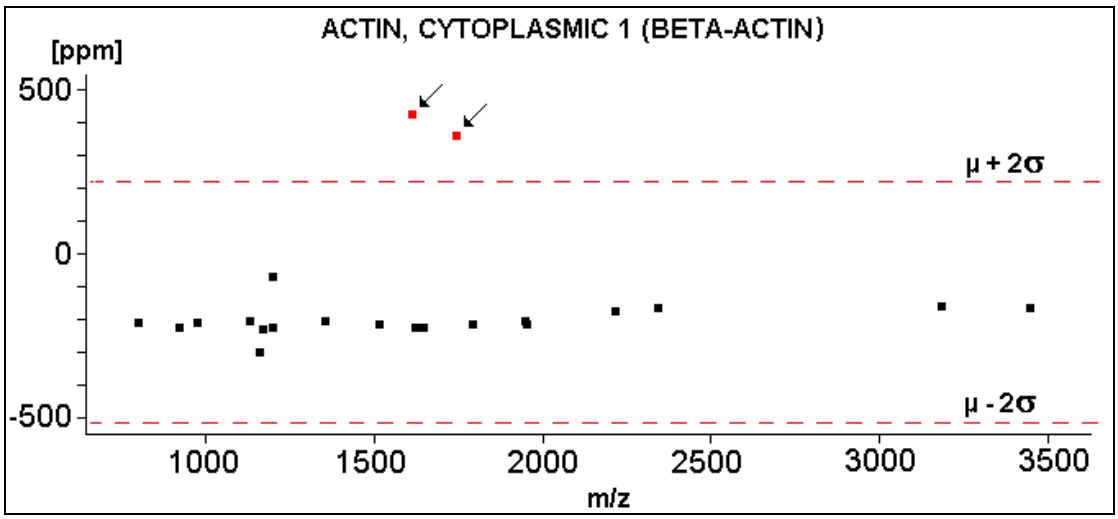

Abbildung 17 **Relative Abweichungen in ppm der experimentell bestimmten Massen von den für das Protein „*Actin, Cytoplasmic 1 (Beta-Actin)*“ berechneten Peptidmassen, die in in einem Fehlerintervall von ± 500 ppm mit diesen übereinstimmen, versus den theoretisch berechneten Peptidmassen.** Die rot gestrichelten Linien zeigen den erlaubten Bereich, in dem sich die Abweichungen bewegen dürfen, um bei der weiteren Berechnung berücksichtigt zu werden. Die mit Pfeilen markierten und rot gefärbten Werte wurden bei diesem Schritt ausgeschlossen.

Die anschließend durchgeführte lineare Regression (siehe Schritt 3) ist in Abbildung 18 gezeigt und gehorcht folgender Geradengleichung:

y = 0,218x - 238,2.

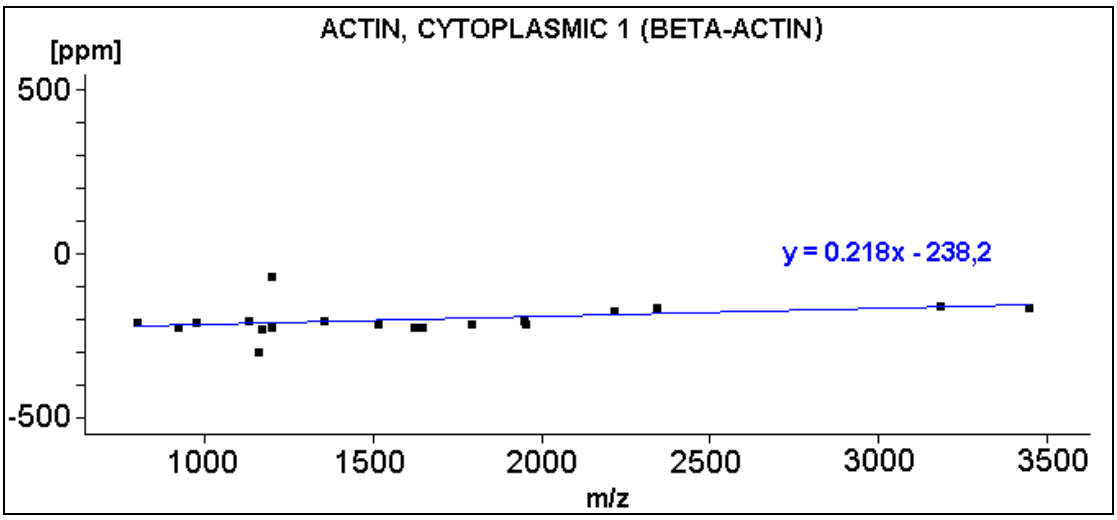

Abbildung 18 Lineare Regression nach der ersten Korrektur (siehe Schritt 2).

Analog zu Schritt 3 wurden die Abweichungen neu berechnet:

a. **y = 0,218 • berechnete Masse –238,2**.

b. neu berechnete Abweichung $\mathbf{ppm_{neu} = ppm_{alt} - y}$.

Aus der neu berechneten Abweichung $\mathbf{ppm_{neu}}$ wurde ein Mittelwert **μ** von $8 \bullet 10^{-7}$ ppm und eine Standardabweichung σ von 40,5 ppm ermittelt. Daraus ergab sich ein erlaubtes Fehlerintervall von **μ-2σ** = –80,9 ppm bis **μ+2σ** = +80,9 ppm. In Abbildung 19 sind die neu berechneten Abweichungen $\mathbf{ppm_{neu}}$ gegen die berechneten Peptidmassen aufgetragen. Bei diesem Schritt wurden zwei weitere Massen herausgefiltert, deren korrigierte Abweichung (-92 ppm bzw. +127 ppm) außerhalb des erlaubten Fehlerintervall liegt (Abbildung 19, mit Pfeilen markiert).

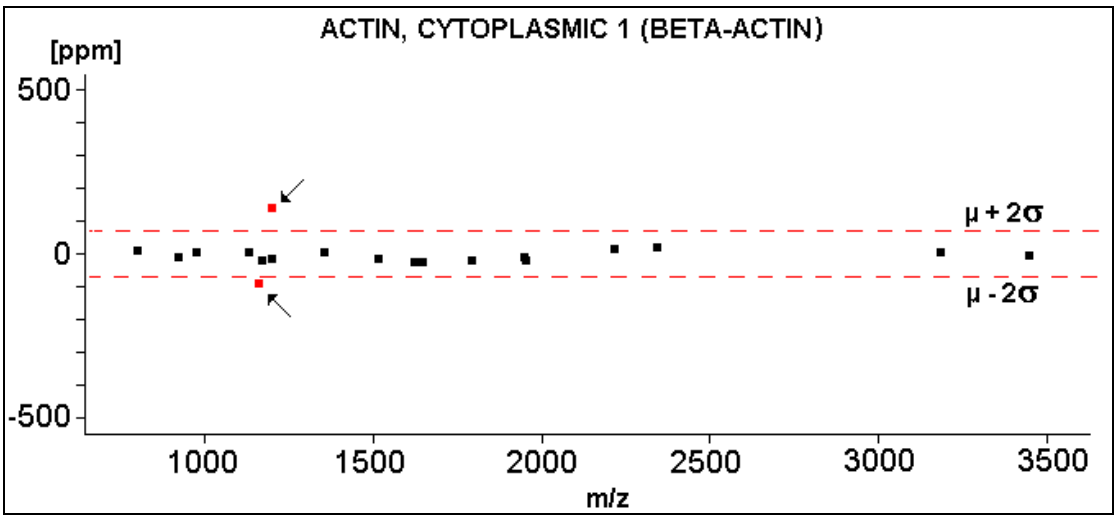

Abbildung 19 Korrigierte relative Abweichungen in ppm versus den berechneten Peptidmassen.
Die rot gestrichelten Linien zeigen den erlaubten Bereich, in dem sich die Abweichungen bewegen dürfen, um bei der weiteren Berechnung berücksichtigt zu werden. Die beiden mit Pfeilen markierten und rot gefärbten Peptidmassen liegen außerhalb dieser Grenzen und wurden somit bei diesem Schritt ausgeschlossen.

Dem untersuchten Protein wurden schließlich 17 von 37 möglichen Massen zugeordnet (siehe Schritt 4). Aus den Abweichungen der verbliebenen Massen wurde eine Standardabweichung σ von 13,7 ppm ermittelt.

Die Anwendung der beschriebenen Algorithmen führt demnach zu einer Reduzierung der Standardabweichung σ von ursprünglich 180,7 ppm auf 13,7 ppm. Die Verbesserung der Massenrichtigkeit erfolgt bei dem beschriebenen Verfahren nicht durch Korrektur der Absolutmassen, sondern durch Selektion derjenigen Peptidmassen die nicht mehr als $\pm$ **2**σ um ihren Mittelwert **μ** streuen. Dadurch ist es möglich, die für eine eindeutige Identifizierung von Proteinen in großen Datenbanken geforderte Spezifität von mindestens 50 ppm, auch ohne Verwendung interner Kalibranten (siehe 1.5.1) zu erreichen. Da die Algorithmen auf alle in der ausgewählten Datenbank befindlichen Proteine angewendet werden, ergeben sich für jedes dieser Proteine mehr oder weniger unterschiedliche Standardabweichungen.

Das richtige Protein zeichnet sich jedoch dadurch aus, dass es eine geringe Standardabweichung bei hoher Trefferzahl besitzt. Unter 4.2 wird die Identifizierung der Proteine ausführlich beschrieben.

Die Linearität der Abweichungen über den gesamten betrachteten Massenbereich ist Voraussetzung für die erfolgreiche Anwendung des oben beschriebenen Verfahrens und ist abhängig von der verwendeten Kalibrierungsfunktion. Bei der im oben beschriebenen Beispiel verwendeten 2-Punkt-Kalibrierung wurden die Flugzeiten von zwei bekannten Peptidmassen (Angiotensin II (human) M_G =1046,5423 Da und ACTH human (CLIP) (18-39) M_G = 2465,1989 Da) bestimmt und die zwei Kalibrierungskonstanten a_0 bzw. a_1 nach Gleichung 7 **a** und **b** ermittelt.

$$\textbf{a:} \quad m/z_{(Ang\ II)} = a_0 + a_1 (t_{(Ang\ II)})^2 \qquad (7)$$

$$\textbf{b:} \quad m/z_{(ACTH)} = a_0 + a_1 (t_{(ACTH)})^2$$

Mit den Kalibrierungskonstanten a_0 und a_1 wurden aus den Flugzeiten die Peptidmassen der Probe ermittelt. Die Verwendung von zwei externen Kalibranten zur Massenbestimmung hat jedoch noch relativ hohe systematische Abweichungen für die zwischen den Kalibranten liegenden m/z-Intervalle zur Folge. In enger Zusammenarbeit mit Dr. J. Gobom, Dr. E. Nordhoff und Dipl.-Ing. M. Müller wurde von mir ein neues Kalibrierungsverfahren entwickelt, mit dem die Linearität der Abweichungen erhöht und somit die Standardabweichung σ auf unter 10 ppm reduziert werden kann [77]. Für eine externe Kalibrierung wurden zunächst die Flugzeiten von 58 bekannten Massen eines Polypropylenglykolgemisches (PPG, siehe 3.2.3), die über einen Massenbereich von 737 bis 4045 m/z gleichmäßig verteilt sind, bestimmt und damit für folgende Polynomfunktion 15^{ten} Grades 16 Kalibrierungskonstanten berechnet:

$$m/z = a_0 + a_1 (t^2)^1 + a_1 (t^2)^2 \ldots + a_n (t^2)^{15} \qquad (8)$$

Mit diesen Konstanten wurden die Massen aus den gemessenen Flugzeiten ermittelt. Abbildung 20 zeigt ein MALDI-Massenspektrum des verwendeten Polypropylenglykolgemisches.

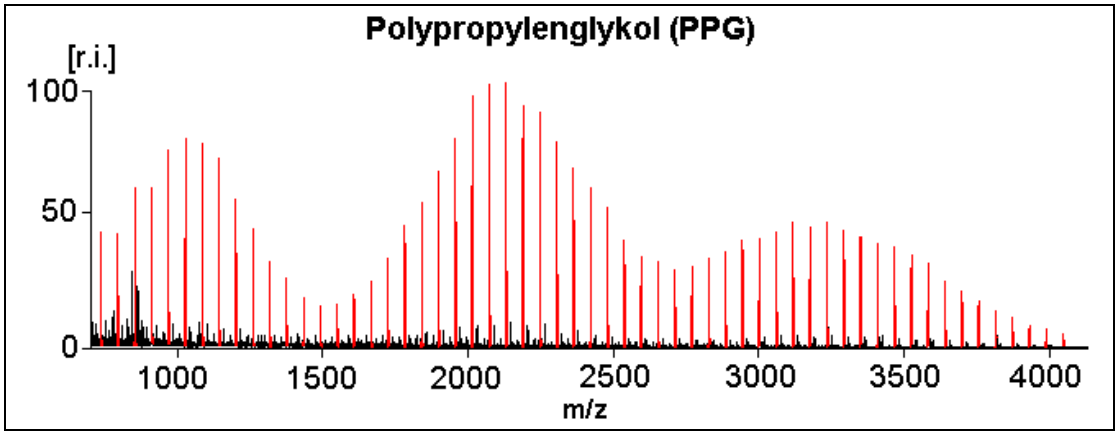

Abbildung 20 MALDI-TOF-Massenspektrum von dem für die Kalibrierung verwendeten Polypropylenglykolgemisches.

In einem Experiment wurde das Polypropylenglykolgemisch (PPG-Standard) an den Position H12, D6, D18, L6 und L18 auf einem 384er MALDI-Probenträgers aufgetragen (Abbildung 21). Die Probenpräparation wurde wie unter 3.2.2, das „*Peakpicking*" wie unter 3.2.1 beschrieben durchgeführt.

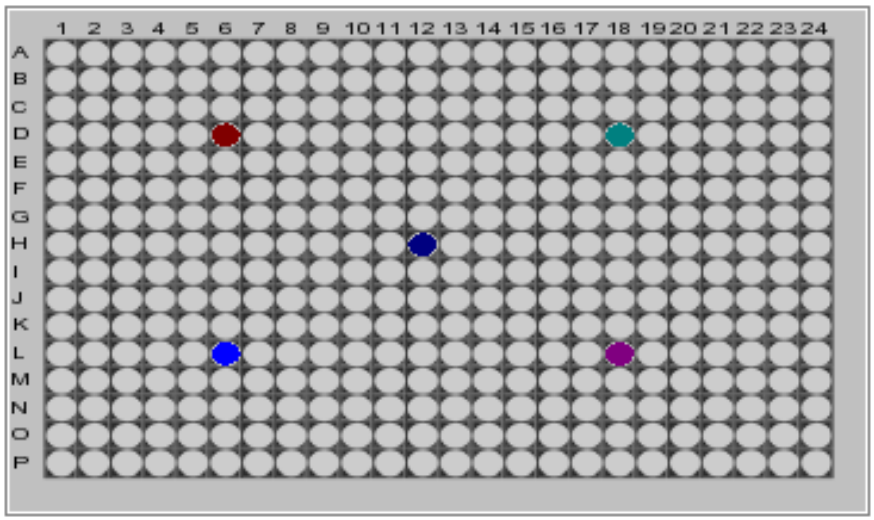

Abbildung 21 Positionierung der präparierten PPG-Proben auf dem verwendeten 384er MALDI-Probenträger.

Die Molekülmassen der verschiedenen Komponenten des PPG-Standards wurden aus den Quadraten der gemessenen Flugzeiten (t^2) einmal mit Hilfe einer 2-Punkt-Kalibrierung (Gleichung 7) und einmal mit Hilfe einer Polynomfunktion 15^{ten} Grades (Gleichung 8) berechnet. In Abbildung 22-A sind für die beiden Verfahren (interne Kalibrierung) die resultierenden relativen Abweichungen der bestimmten von den berechneten richtigen Werten gegenübergestellt. Für die 2-Punkt-Kalibrierung wurden m/z 737,5027 und 4045,889 des PPG-Standards als interne Kalibranten verwendet. Für die Berechnung der Konstanten der Polynomfunktion 15^{ten} Grades wurden alle monoisotopischen m/z-Werte des PPG-Standards verwendet. Im Fall der 2-Punkt-Kalibrierung weichen die bestimmten Massen von ihren Sollwerten um bis zu maximal 50 ppm ab. Der Verlauf der Abweichung in Abhängigkeit vom Sollwert zeigt deutlich, dass die beobachteten Fehler im wesentlichen systematischer sind. Es folgt dass für die nötige Umrechnung eine lineare Kalibrierungsfunktion nicht geeignet ist. Wie erwartet, liefert eine Vielpunkt-Kalibrierung kombiniert mit einer Polynomfunktion mit einer hinreichenden Zahl von Korrekturgliedern bessere Ergebnisse.

Die mit Hilfe des von der Position H12 aufgenommenen Spektrums bestimmten Kalibrierkonstanten wurden anschließend verwendet um die für die Positionen D6, D18, L6 und L18 bestimmten Werte t^2 aller Komponenten des PPG-Standards in die entsprechenden m/z-Werte umzurechnen (externe Kalibrierung). Die Ergebnisse sind in Abbildung 22-B und Abbildung 22-C für die beiden genannten Kalibrierverfahren gegenübergestellt.

In beiden Fällen (externe Kalibrierung, Abbildung 22-B und C) schwankt die relative Abweichung der bestimmten von den korrekten Werten je nach Probenposition von +20 ppm bis –130 ppm. Es ist deutlich zu ersehen, insbesondere wenn für die Umrechnung von t^2 in m/z die für H12 optimierte Polynomfunktion 15^{ten} Grades verwendet wird, dass die mit einem Positionswechsel verbundenen Fehler systematisch sind und für die verschiedenen Positionen in erster Näherung mit eine linearen Tranformationsgleichung beschrieben werden können.

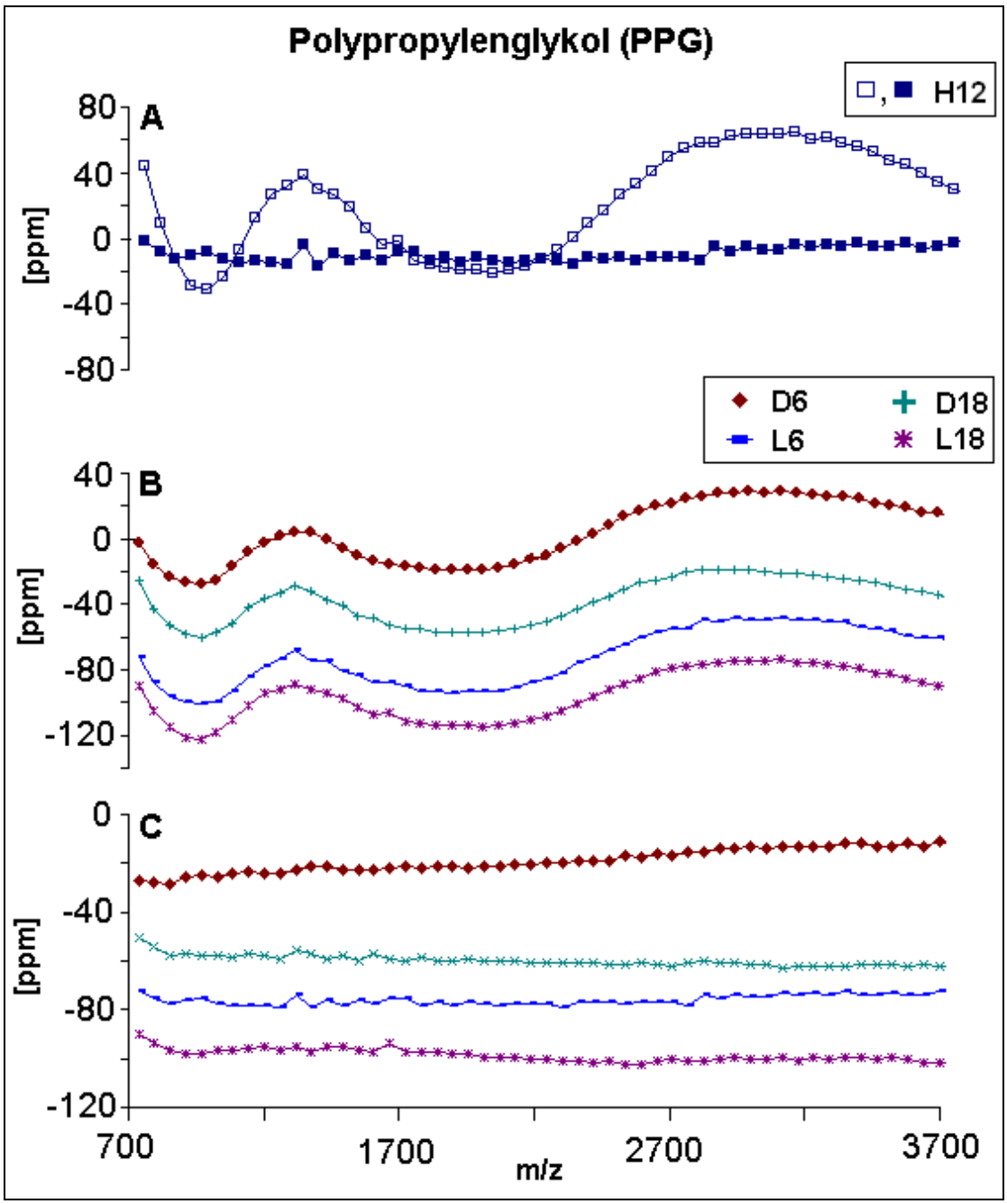

Abbildung 22 (A) Vergleich der relativen Abweichungen der aus dem Quadrat aller für den PPG-Standard auf Position H12 bestimmten Flugzeiten berechneten Molekülmassen von den richtigen Werten nach interner 2-Punkt-Kalibrierung (m/z 737,5027 und m/z 4045,8890, offene Quadrate) und alternativ, nach Umrechnung mit Hilfe einer für den Datensatz optimierten Polynomfunktion 15ten Grades (gefüllte Quadrate). (B) Die für die Position H12 etablierte lineare Kalibrierungsfunktion wurde anschließend auf die Positionen D6, D18, L6 und L18 angewendet. Die Gegenüberstellung zeigt die mit dem Positionswechsel verbundene Änderung der relativen Abweichungen. (C) Analog zu (B) mit dem Unterschied dass für die Berechnung der Molekülmassen die für Position H12 optimierte Polynomfunktion verwendet wurde.

Im folgenden wird die Effizienz der verbesserten Kalibrierungsfunktion an einem konkreten Beispiel veranschaulicht. Eine Probe des rekombinierten menschlichen Proteins „*MRNA-Associated Portein MRNP 41*“ wurde an der Position G16, der für die Kalibrierung verwendete PPG-Standard (siehe 3.2.3) wurde in unmittelbarer Nähe an der Position G12 des verwendeten 384er MALDI-Probenträgers aufgetragen (nicht gezeigt). Die Probenpräparation wurde wie unter 3.2.2, das „*Peakpicking“* wie unter 3.2.1 beschrieben durchgeführt. Die Massen der Probe wurden einmal mit Hilfe einer 2-Punkt-Kalibrierung (Gleichung 7) und einmal mit Hilfe einer Polynomfunktion 15ten Grades berechneten Kalibrierung (Gleichung 8) aus den gemessenen Flugzeiten des PPG-Standards berechnet. Die resultierenden Peaklisten enthielten 48 Peaks, die im Massenspektrum des Proteins (Abbildung 23) grün bzw. rot eingefärbt sind. Grün gefärbte Peaks entsprechen Peptiden des gesuchten Proteins, rot sind Peaks die nicht zugeordnet werden konnten.

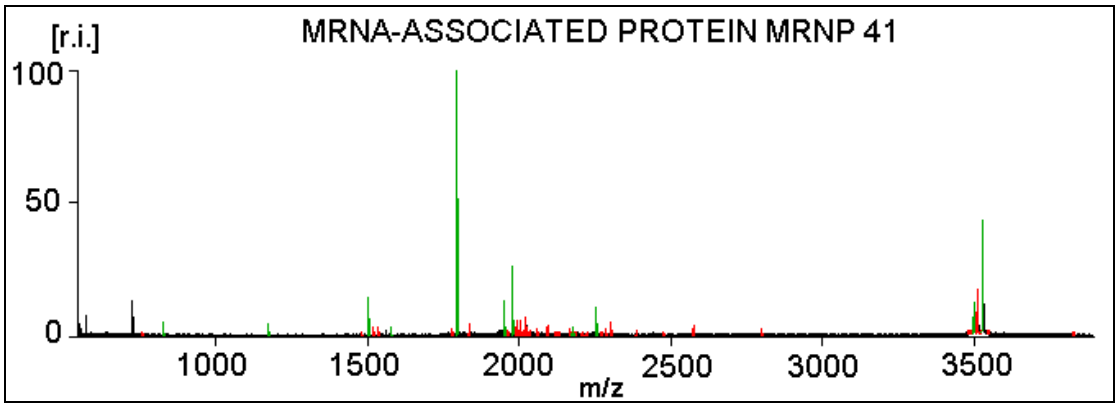

Abbildung 23 MALDI-TOF-Massenspektrum des tryptisch verdauten rekombinierten menschlichen Protein „MRNA-Associated Portein MRNP 41“.
Grün ein-gefärbte Peaks symbolisieren Peptide des Proteins, rote Peaks konnten nicht zugeordnet werden.

Nach Anwendung der oben beschrieben Algorithmen (Schritt 1-4) wurden dem Protein „*MRNA-Associated Portein MRNP 41“* 11 Massen zugeordnet.
In Abbildung 24 sind die Abweichungen in ppm der experimentell bestimmten Massen von den für das untersuchte Protein theoretisch berechneten Peptidmassen versus den theoretisch berechneten Peptidmassen aufgetragen.

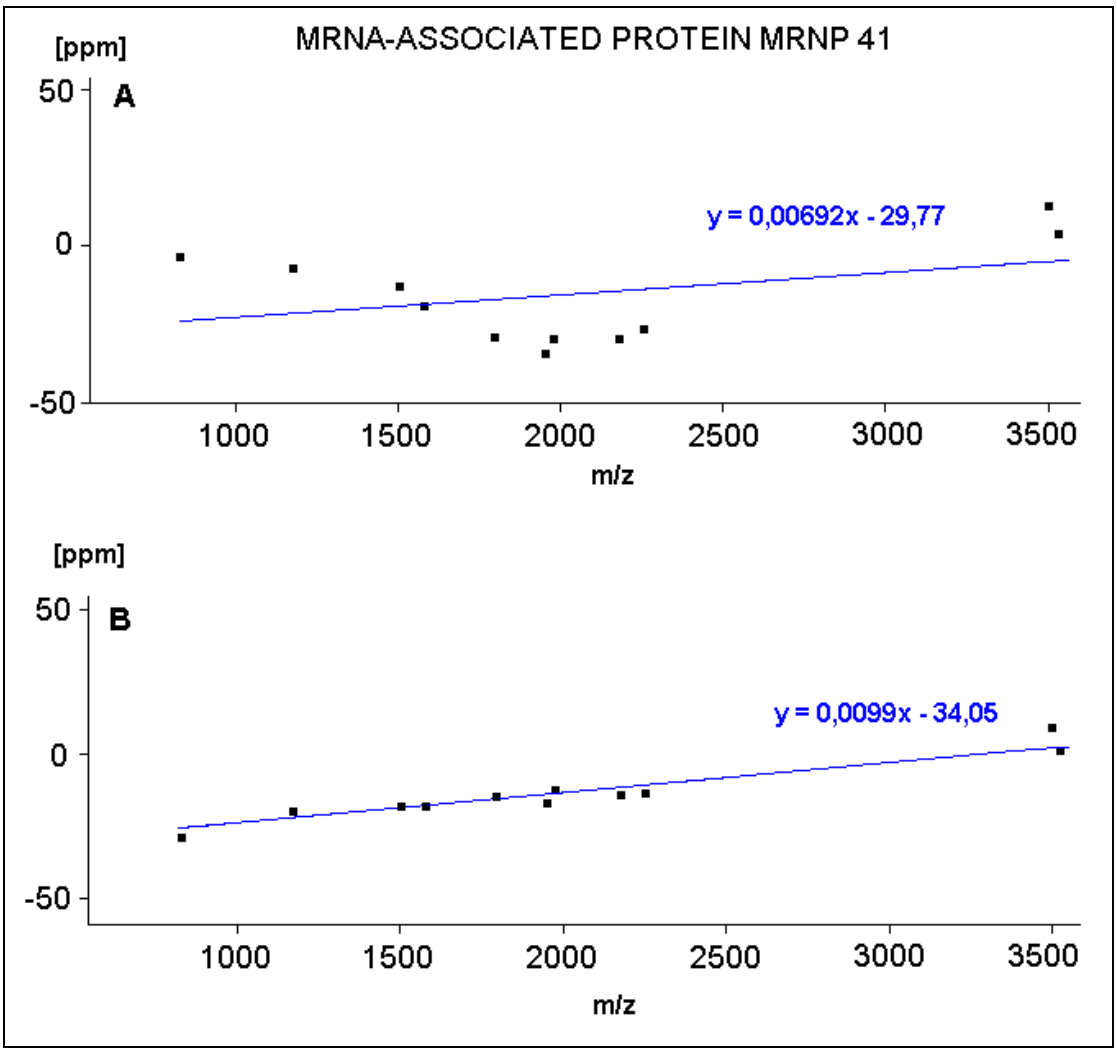

Abbildung 24 Vergleich von 2-Punkt-Kalibrierung (A) und einer mit Hilfe einer Polynomfunktion 15ten Grades berechneten Kalibrierung (B).
Dargestelt sind die korrigierten relativen Abweichungen in ppm versus den theoretisch berechneten Peptidmassen.

Bei Verwendung der 2-Punkt-Kalibrierung (Abbildung 24-A) wurde eine Standardabweichung σ von **13,9 ppm** ermittelt, bei mit der Polynomfunktion 15^{ten} Grades berechneten Kalibrierung (Abbildung 24-B) betrug die Standardabweichung σ nur noch **1,9 ppm**. Mit Hilfe der durch die extern bestimmten Polynomfunktion 15^{ten} Grades berechneten m/z-Werte wird die Streuung um 12 ppm gegenüber der 2-Punkt-Kalibrierung gesenkt. Daraus folgt, dass die Kombination aus mit der extern bestimmten Polynomfunktion 15^{ten} Grades berechneten m/z-Werte (Linearisiierung der Abweichungen um

die durch lineare Regression berechnete Gerade) und den oben beschriebenen Algorithmen (Linearisierung der Abweichungen um die durch lineare Regression berechnete Gerade und Berücksichtigung der Nullpunktverschiebung) das Mittel der Wahl ist, um bei externer Kalibrierung die geforderte Spezifität von mindestens 50 ppm zur eindeutigen Identifizierung eines Proteins in großen Datenbanken zu unterschreiten.

4.2 „*Scoring*"-Algorithmus zur sichereren Identifikation von Proteinen in großen Sequenzdatenbanken

Wie unter 1.5.3 bereits beschrieben, liefern etablierte „*Scoring*"-Algorithmen nur zuverlässige Resultate, wenn die Massenrichtigkeit besser als 50 ppm ist. Der im folgenden von mir entwickelte Algorithmus ermittelt aus den Parametern Standardabweichung σ, Trefferanzahl („Hits" siehe 4.1, Schritt 4) und prozentualen Sequenzabdeckung der Proteine **SC** (siehe 4.1, Schritt 5) für jedes Protein einen „*Scoring*"-Faktor **Z**. Dieser wird nach folgender Formel berechnet:

$$Z = 100 - \frac{F \times 500 \times \sigma}{Hits^2 \times SC} \qquad \text{Für} \quad \begin{matrix} Hits > 4 \\ SC > 0 \end{matrix} \qquad \textbf{(9)}$$

Es werden nur Proteine berücksichtigt, die mehr als 4 Treffer liefern und deren **SC**-Wert > 0 ist.

F ist ein dimensionsloser Faktor, der je nach Anforderung angepasst werden kann. Im Falle einer großen Sequenzdatenbank (z.B. NCBI „*non redundant*") wird F auf 1 gesetzt, wird hingegen nur eine relativ kleine Sequenzdatenbank (z.B Alle Einträge von E.Coli in der Swiss-Prot) kann im F entsprechend angepaßt werden (z.B. F=0.5). **Z** kann Werte zwischen 0 und 100 annehmen.

Je mehr Treffer ein Protein liefert, desto geringer muß die Standard-abweichung sein und umgekehrt je niedriger die Standardabweichung, desto weniger Treffer sind nötig um einen akzeptablen Wert für **Z** zu erhalten. Da große Proteine eine Vielzahl von Peptiden liefern, ist es möglich, dass einige dieser Peptide zufällig Abweichungen zu ihren korrespondierenden experimentell bestimmten Massen besitzen, die so verteilt sind, dass die resultierende Standardabweichung σ sehr niedrig ist und diese fälschlicherweise als möglicher Kandidat Berücksichtigung finden könnten.
Der **SC**-Wert verhindert dies, da große Proteine auch eine entsprechend größere Summe an Peptidsequenzen $L_{Peptide}$ (siehe Gleichung 6) aufweisen müssen, um einen ausreichenden Wert für **SC** zu erreichen.

Die Berechnung von Z wurde so gewählt, dass gilt:

- Ein Protein ist sicher identifiziert, wenn $\mathbf{Z} \geq 99$.
- Wenn $\mathbf{Z} < 99$ und $\mathbf{Z} \geq 98$ handelt sich um einen möglichen Kandidaten. Für eine sichere Identifizierung werden aber zusätzliche Informationen benötigt.
- Wenn $\mathbf{Z} < 98$ ist das gefundene nicht mit dem gesuchten Protein identisch.

Im folgenden wird die Effizienz des „*Scoring*“-Algorithmus an zwei Beispielen demonstriert. Es wurden je eine Probe der tryptischen Spaltprodukte der gentechnisch hergestellten menschlichen Proteine „*Heatshock Cognate 71Kd Protein*“ im folgenden als HSP bezeichnet (Position C19, Abbildung 25) und „*60S Ribosomal Protein L7*“ im folgenden als 60SRP bezeichnet (Position L5, Abbildung 25) auf einem 384er MALDI-Probenträger aufgetragen.

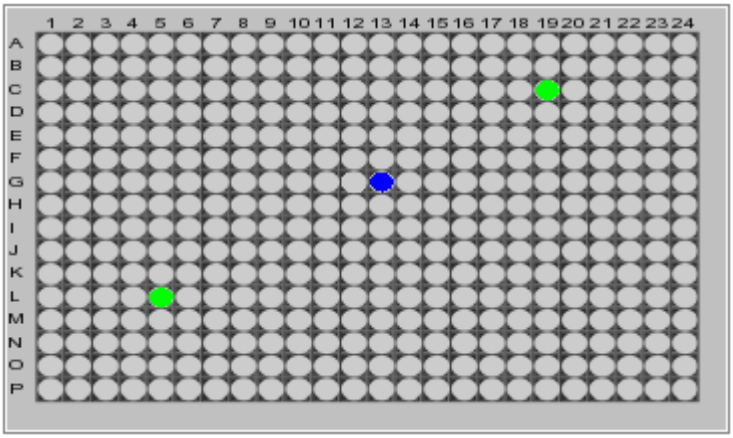

Abbildung 25 Schematische Darstellung des 384er MALDI-Probenträgers.
Der PPG-Standard wurde an Position G13 (blau markiert) aufgetragen. Die Probe der tryptischen Spaltprodukte des Proteins „*Heatshock Cognate 71Kd Protein*“ befindet sich an Position C19 (grün markiert) und die des Proteins „*60S Ribosomal Protein L7*“ an Position L5 (ebenfalls grün markiert).

Der für die Kalibrierung verwendete PPG-Standard (siehe 3.2.3) wurde in der Mitte des MALDI-Probenträgers an der Position G13 aufgetragen (Abbildung 25). Die Probenpräparation wurde wie unter 3.2.2, das „*Peakpicking*“ wie unter 3.2.1 beschrieben durchgeführt. Die Massen wurden mit Hilfe einer Polynomfunktion 15^{ten} Grades aus den bestimmten Flugzeiten für die die 16 Konstanten a_0-a_{15} mit Hilfe des PPG-Standards bestimmt wurden, berechnet (siehe 4.3). Die resultierenden Peaklisten enthielten 115 Einträge für HSP und 92 für 60SRP. Bei diesem und allen folgenden Beispielen wurde, wenn nicht anders beschrieben für die Datenbanksuche die Swiss-Prot-Datenbanken *für „Mus musculus*“ und „*Homo Sapiens*“ ausgewählt, der dimensionslose Faktor F auf 1 und der zu erwartende maximale Fehler auf 500 pmm gesetzt. Mögliche Modifikationen blieben unberücksichtigt und alle Proteine der ausgewählten Datenbanken wurden unabhängig von ihrer Größe berücksichtigt.

4.2.1 Beispiel für die Effizienz des „Scoring“-Algorithmus bei hoher Trefferzahl

Abbildung 26 zeigt das Massenspektrum der von HSP gewonnen tryptischen Spaltprodukte.

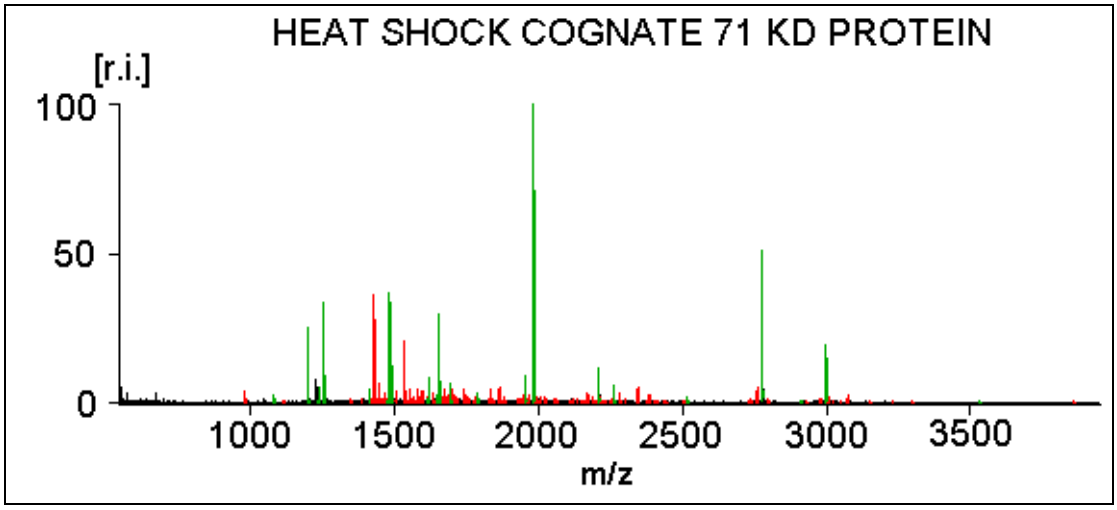

Abbildung 26 **MALDI-TOF-Massenspektrum des tryptisch verdauten rekombinierten menschlichen Protein „*Heatshock Cognate 71Kd Protein*“(HSP).**
Grün eingefärbte Peaks wurden HSP zugeordnet, rote Peaks nicht.

Die ersten fünf Kandidaten der aus der Datenbanksuche für HSP generierten „*Scoring*“-Liste sind in Tabelle 6 aufgelistet.

Tabelle 6 **Ergebnis der Datenbanksuche für „*Heatshock Cognate 71Kd Protein*“ (HSP). (M) steht für Mus musculus und (H) für Homo sapiens.**

Rang	Protein	Swiss-Prot Zugriffs-Nr.	MG [Da]	Z	Hits	σ [ppm]	SC [%]
1	HEAT SHOCK COGNATE 71 KD PROTEIN. **(H)**	P11142	70,898	99.8	22	9.5	58.7
2	HEAT SHOCK COGNATE 71 KD PROTEIN. **(M)**	P08109,P12225, Q62373,Q62374	70,871	99.8	22	9.5	58.7
3	MYOSIN HEAVY CHAIN, CARDIAC MUSCLE BETA ISOFORM. **(H)**	P12883,Q14904, Q16579	223,112	98.8	31	54.7	24.3
4	MYOSIN HEAVY CHAIN, CARDIAC MUSCLE ALPHA ISOFORM. **(M)**	Q02566,Q64258, Q64738	223,564	98.2	27	54.3	20.3
5	ENDOPLASMIN PRECURSOR **(H)**	P14625	92,468	97.1	12	21	25.3

Das gesuchte Protein wurde durch den Scoring-Algorithmus mit einem Z-Faktor von 99,8 richtig identifiziert (Rang1, Tabelle 6). An Rang 2 ist das homologe Protein aus „Mus musculus“ mit dem gleichen Z-Faktor gelistet. Eine Differenzierung zwischen diesen beiden Proteinen ist nicht möglich, da sie sich in ihrer Sequenz lediglich durch eine einzige Aminosäure unterscheiden (an Position 579 der Aminosäuresequenz im HSP aus „Mus musculus“ ist Asparagin durch Serin ersetzt; in Abbildung 27 rot markiert) und diese zu keiner der bei der Suche gefundenen Peptidsequenzen (in Abbildung 27 grün gefärbt), deren Masse mit einer der gesuchten Massen übereinstimmt, gehört. Die Suche ergab für beide Proteine je 22 Treffer, je eine Standardabweichung von 9,5 ppm und je einen SC-Wert von 58,7 %. In diesem Fall sind beide Proteine als gleich wahrscheinliche Kandidaten anzusehen und werden gleichberechtigt durch den Algorithmus als richtige Kandidaten vorgeschlagen. Die korrekte Identifizierung der Spezies ist in diesem Fall nicht möglich, da die hierfür nötige Information (siehe oben und Abbildung 27) in den experimentell generierten Daten nicht enthalten ist. Ist die Spezies jedoch vorab bekannr, so kann diese Information vorab in der Datenbanksuche mit berücksichtigt werden (siehe Abbildung 33).Das an Rang 3 gelistete *Protein „Myosin Heavy Chain, Cardiac Muscle Beta Isoform*“ besitzt zwar 9 Treffer mehr als die richtigen Kandidaten, hat aber aufgrund der relativ hohen Standardabweichung von 54,7 ppm und dem vergleichsweise niedrigen SC-Wert von 24,3 nur einen Z-Faktor von 98,8 und bleibt damit unter dem für eine eindeutige Identifizierung geforderten Wert von $Z \geq 99$.

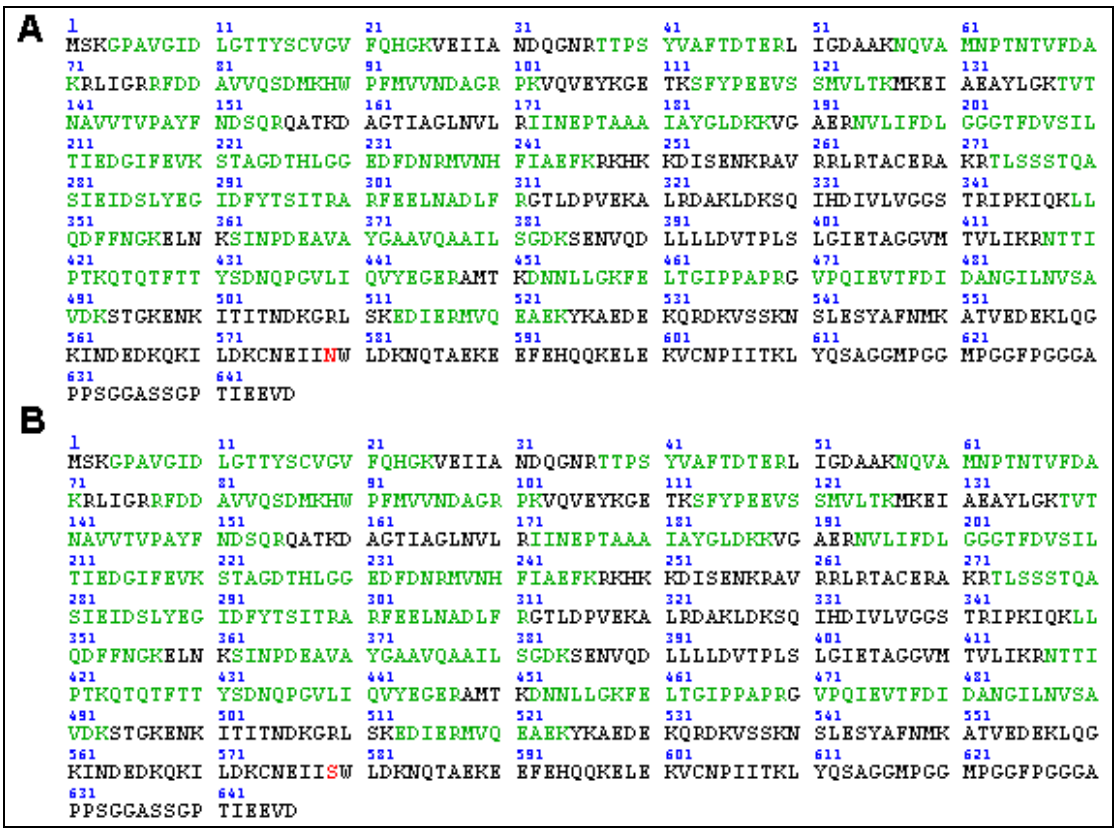

Abbildung 27 Aminosäuresequenz von „*Heatshock Cognate 71Kd Protein*" aus "*Homo Sapiens*" (A) und "*Mus musculus*" (B).
Beide Sequenzen unterscheiden sich in einer einzigen Aminosäure an Position 579 (rot gefärbt). Grün gefärbte Abschnitte kennzeichnen Peptidsequenzen, deren Massen experimentell bestimmten Peptidmassen zugeordnet wurden.

Wie unter 4.1 bereits erwähnt, zeichnet sich das richtige Protein dadurch aus, dass es eine geringe Standardabweichung bei hoher Trefferzahl besitzt. Abbildung 28 verdeutlicht dies. Eine hohe Trefferzahl oder eine niedrige Standardabweichung als alleiniges Kriterium reicht jedoch nicht aus, um eine eindeutige Identifizierung zu gewährleisten. Um dies zu verdeutlichen wurde die Datenbanksuche mit der gleichen Probe und unter gleichen Bedingungen zweimal wiederholt. Zunächst mit der Einschränkung, dass die gefundenen Proteine nach ihrer Standardabweichung aufsteigend sortiert (Tabelle 7) und anschließend mit der Einschränkung, dass die gefundenen Proteine nach ihrer Trefferanzahl absteigend sortiert (Tabelle 8) wurden.

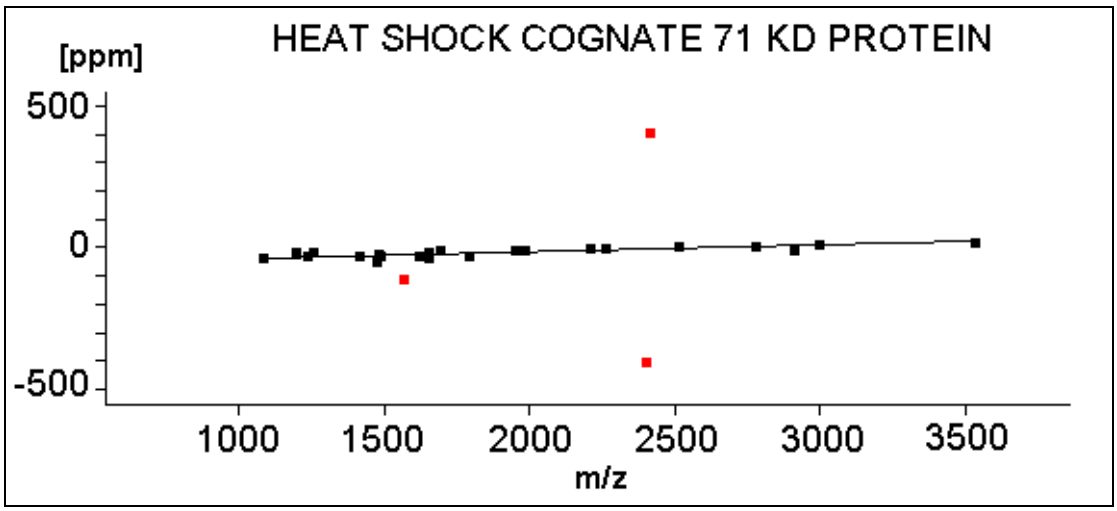

Abbildung 28 **Darstellung der relativen Abweichungen in ppm der experimentell bestimmten versus den berechneten Peptidmassen des Proteins „*Heatshock Cognate 71Kd Protein*“(HSP).**
Die rot eingefärbten Peptidmassen sind Ausreißer die herausgefiltert wurden.

In Tabelle 7 ist das *Protein „Rac-Alpha Serine/Threonine Kinase*“ aus „*Homo Sapiens*“ mit der niedrigsten Standardabweichung aller Proteine von σ = 3,4 ppm aber mit nur 5 Treffern auf Rang 1 gelistet. Das richtige Protein folgt erst auf Rang 9 (Tabelle 7, grüner Hintergrund). In Tabelle 8 hingegen, ist das Protein *„Myosin Heavy Chain, Cardiac Muscle beta Isoform“* mit 31 Treffern und einer hohen Standardabweichung von 54,7 ppm auf Rang 1 gelistet. Das richtige Protein folgt hier auf Rang 4 (Tabelle 8, grüner Hintergrund). In beiden Fällen wurde das richtige Protein nicht identifiziert.

Tabelle 7 **Ergebnis der Datenbanksuche für „*Heatshock Cognate 71Kd Protein*“ (HSP) nach Standardabweichung aufsteigend sortiert. (M) steht für Mus musculus und (H) für Homo sapiens.**

Rang	Protein	Swiss-Prot Zugriffs-Nr.	MG [Da]	Z	Hits	σ [ppm]	SC [%]
1	RAC-ALPHA SERINE/THREONINE KINASE (EC 2.7.1.-) **(M)**	P31750	55,622	96.4	5	3.4	18.8
2	2',3'-CYCLIC NUCLEOTIDE 3'-PHOSPHODIESTERASE (EC 3.1.4.37) (CNP) **(M)**	P16330	47,123	94.5	5	4.9	17.9
3	CATALASE (EC 1.11.1.6). **(H)**	P04040	59,756	92.9	5	5.5	15.6

Rang	Protein	Swiss-Prot Zugriffs-Nr.	MG [Da]	Z	Hits	σ [ppm]	SC [%]
4	SODIUM/GLUCOSE CO-TRANSPORTER 1 (NA (+)/GLUCOSE COTRANSPORTER 1) **(H)**	P13866	73,497	87.4	5	7.7	12.2
5	MITOCHONDRIAL TRANSCRIPTION FACTOR 1 PRECURSOR (MTTF1).**(H)**	Q00059	29,096	94.6	5	7.8	29.3
	• • •						
9	HEAT SHOCK COGNATE 71 KD PROTEIN. **(H)**	P11142	70,898	99.8	22	9.5	58.7

Tabelle 8 **Ergebnis der Datenbanksuche für „*Heatshock Cognate 71Kd Protein*“ (HSP) nach Trefferanzahl absteigend sortiert. (M) steht für Mus musculus und (H) für Homo sapiens.**

Rang	Protein	Swiss-Prot Zugriffs-Nr.	MG [Da]	Z	Hits	σ [ppm]	SC [%]
1	MYOSIN HEAVY CHAIN, CARDIAC MUSCLE BETA ISOFORM. **(H)**	P12883 Q14904 Q16579	223,112	98.8	31	54.7	24.3
2	MYOSIN HEAVY CHAIN, CARDIAC MUSCLE ALPHA ISOFORM. **(H)**	P13533 Q13943 Q14906 Q14907	223,689	97	27	92	21.4
3	MYOSIN HEAVY CHAIN, CARDIAC MUSCLE ALPHA ISOFORM. **(M)**	Q02566 Q64258 Q64738	223,564	98.2	27	54.3	20.3
4	HEAT SHOCK COGNATE 71 KD PROTEIN. **(H)**	P11142	70,898	99.8	22	9.5	58.7
5	HEAT SHOCK COGNATE 71 KD PROTEIN. **(M)**	P08109 P12225 Q62373 Q62374 Q62375	70,871	99.8	22	9.5	58.7

Das Protein *„Myosin Heavy Chain, Cardiac Muscle beta Isoform“* besitzt zwar mit Abstand die höchste Trefferzahl, jedoch ist die Streuung der für dieses Protein gefundenen 31 Peptidmassen um die durch lineare Regression berechneten Geraden zu hoch (Abbildung 29).

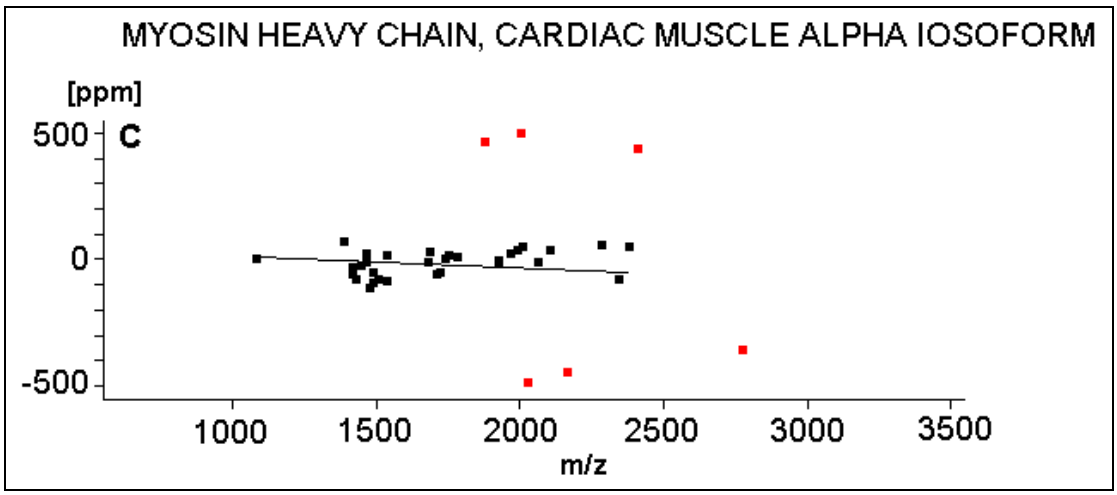

Abbildung 29 Darstellung der relativen Abweichungen der experimentell bestimmten versus den berechneten Peptidmassen des Proteins „*Myosin Heavy Chain, Cardiac Muscle beta Isoform*“ in ppm.
Die rot eingefärbten Peptidmassen sind Ausreißer die herausgefiltert wurden.

4.2.2 Beispiel für die Effizienz des „Scoring“-Algorithmus bei niedriger Trefferzahl

Abbildung 30 zeigt das Massenspektrum der von 60SRP gewonnen tryptischen Spaltprodukte.

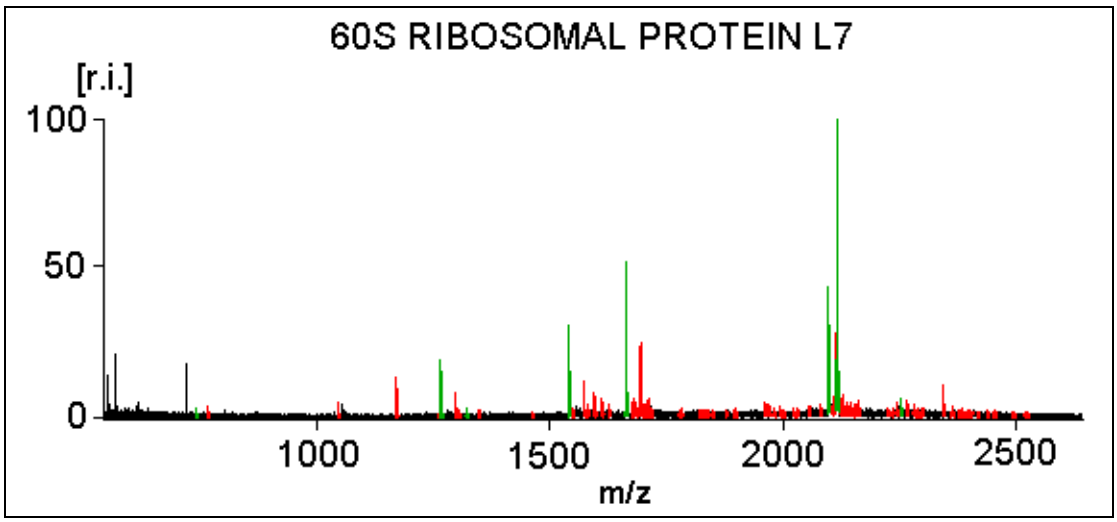

Abbildung 30 MALDI-TOF-Massenspektrum des tryptisch verdauten rekombinierten menschlichen Proteins „*60S Ribosomal Protein L7*“ (60SRP).
Grün einge-färbte Peaks wurden 60SRP zugeordnet, rote Peaks nicht.

Die ersten fünf Kandidaten der aus der Datenbanksuche für 60SRP generierten „*Scoring*“-Liste sind in Tabelle 9 aufgelistet.

Tabelle 9 **Ergebnis der Datenbanksuche für „60S Ribosomal Protein L7“(60SRP). (M) steht für Mus musculus und (H) für Homo sapiens.**

Rang	Protein	Swiss-Prot Zugriffs-Nr.	MG [Da]	Z	Hits	σ [ppm]	SC [%]
1	60S RIBOSOMAL PROTEIN L7. **(H)**	P18124	29,226	99	8	5.7	44.4
2	M-PHASE INDUCER PHOSPHATASE 1 (EC 3.1.3.48) **(H)**	P30304	58,796	98.1	5	1.2	13.4
3	60S RIBOSOMAL PROTEIN L7. **(M)**	P14148	31,419	97.1	6	5.6	27
4	LAMINS C AND C2. **(M)**	P11516	65,446	96	5	2.9	14.6
5	LAMIN A. **(M)**	P48678 P97859	74,209	95.4	5	2.9	12.6

Das gesuchte Protein wurde durch den Scoring-Algorithmus mit einem Z-Faktor von 99,0 richtig identifiziert (Rang1, Tabelle 9). An Rang 2 ist das *Protein „M-Phase Inducer Phosphatase 1 (EC 3.1.3.48)*” mit Z=98,1 bei 5 Treffern, einer sehr niedrigen Standardabweichung von 1,2 ppm und einem vergleichsweißen niedrigen SC-Wert gelistet.

Wie oben bereits erwähnt, können größere Proteine einige Peptide liefern, deren Abweichungen zu ihren korrespondierenden gemessenen Massen so verteilt sind, dass die resultierende Standardabweichung sehr niedrig ist und sie somit fälschlicherweise als möglicher Kandidat Berücksichtigung finden.

Um dies zu verdeutlichen wurde die Datenbanksuche für 60SRP unter den gleichen Bedingungen wiederholt, mit der Einschränkung, dass der SC-Wert für alle gefundenen Proteine auf 1 gesetzt wurde. In Tabelle 10 sind die Ergebnisse dieser Suche aufgelistet. Die Ergebnisse zeigen, dass nun das Protein *„M-Phase Inducer Phosphatase 1 (EC 3.1.3.48)*” sich an Rang 1 befindet, das richtige Protein ist auf Rang 2 gelistet. Das erste Protein besitzt 5 Treffer und ist mit einem Molekulargewicht von 58,796 Da annähernd doppelt so groß wie das richtige Protein mit 8 Treffern und einem Molekular-gewicht von 29,226 Da.

Tabelle 10 **Ergebnis der Datenbanksuche für „60S Ribosomal Protein L7“(60SRP). (M) steht für Mus musculus und (H) für Homo sapiens.**

Rang	Protein	Swiss-Prot Zugriffs-Nr.	MG [Da]	Z	Hits	σ [ppm]	SC [%]
1	M-PHASE INDUCER PHOSPHATASE 1 (EC 3.1.3.48) **(H)**	P30304	58,796	75.1	5	1.2	1
2	60S RIBOSOMAL PROTEIN L7. **(H)**	P18124	29,226	55.7	8	5.7	1
3	LAMINS C AND C2. **(M)**	P11516	65,446	42.2	5	2.9	1
4	LAMIN A. **(H)**	P48678 P97859	74,209	42.2	5	2.9	1
5	258.1 KD PROTEIN C21ORF5 (KIAA0933). **(H)**	Q9Y3R5	258,142	27.2	12	21	1

Die Anzahl der möglichen tryptischen Spaltpeptide eines Proteins nimmt mit steigendem Molekulargewicht zu und dadurch auch die Wahrscheinlichkeit, dass die Abweichungen einiger dieser Peptide zu ihren korrespondierenden experimentell bestimmten Massen so verteilt sind, dass deren Streuung um die durch lineare Regression berechnete Gerade geringer ist (Abbildung 31-A), als dies bei dem richtigen Protein der Fall ist (Abbildung 31-B).
Durch Berücksichtigung des SC-Wertes wird dies korrigiert und das richtige Protein identifiziert (Tabelle 9).

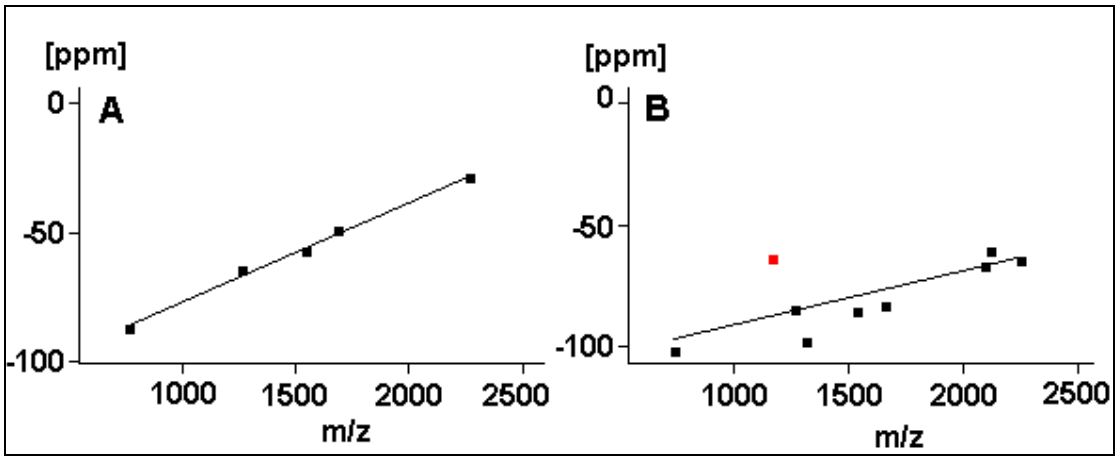

Abbildung 31 **Darstellung der relativen Abweichungen der experimentell bestimmten versus den berechneten Peptidmassen des Proteins „M-Phase Inducer Phosphatase (EC 3.1.1.48)“ in ppm. (A) im Vergleich zu dem Protein „*60S Ribosomal Protein L7*“(B).**
Die rot eingefärbten Peptidmassen sind Ausreißer die herausgefiltert wurden.

4.2.3 Beispiel für die Effizienz des „Scoring“-Algorithmus bei der Suche in großen Datenbanken

Es wurden 96 gentechnisch hergestellte menschliche Proteine mit Trypsin gespalten und jeweils eine probe der generierten Spaltpeptide auf einem 384er MALDI-Probenträger aufgetragen (Abbildung 32). Der für die Kalibrierung verwendete PPG-Standard (siehe 3.2.3) wurde in der Mitte des MALDI-Probenträgers an der Position H12 aufgetragen (Abbildung 32).

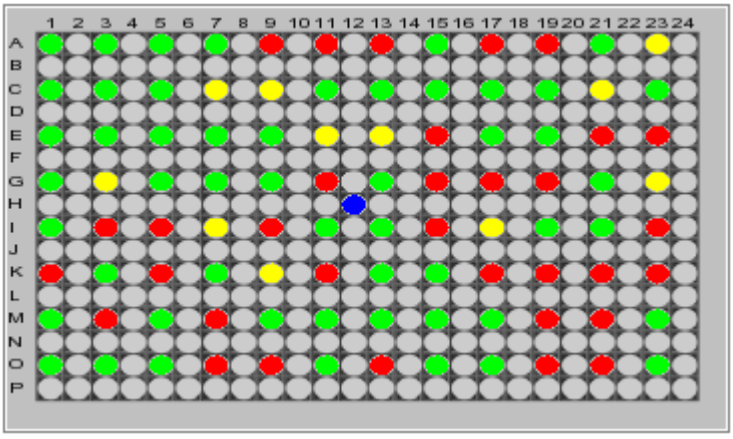

Abbildung 32 Schematische Darstellung des 384er MALDI-Probenträgers.
Der PPG-Standard wurde an Position H12 (blau markiert) aufgetragen. Die Probenpositionen sind entweder grün, gelb oder rot markiert. Grün symbolisiert, dass das entsprechende Protein eindeutig identifiziert wurde ($Z \geq 99$), gelb ($Z \geq 98$ und $Z < 99$) ist nicht sicher identifiziert und rot ($Z < 98$) nicht identifiziert.

Die Probenpräparation wurde wie unter 3.2.2, das „*Peakpicking*“ wie unter 3.2.1 beschrieben durchgeführt. Die Massen wurden mit Hilfe einer Polynomfunktion 15^{ten} Grades für die die 16 Konstanten $a_0 – a_{15}$ mit Hilfe des PPG-Standards bestimmt wurden, berechnet (siehe 4.3). Die Datenbank-suche erfolgte unter den gleichen, oben bereits beschriebenen Bedingungen. Die Suche umfasste alle 686213 in der „*NCBI* non-redundant“ Datenbank eingetragen Proteinsequenzen. Die Ergebnisse sind in Tabelle 11 aufgelistet. In der ersten Spalte stehen die Probenpositionen analog zur Abbildung 32. In der zweiten Spalte sind die Datenbank-Zugriffsnummern und in der dritten

Spalte die Namen der gesuchten Proteine eingetragen. Es wurden 52 von 96 möglichen Proteinen eindeutig identifiziert ($Z \geq 99$). Die entsprechenden Proteine sind in Tabelle 11 grün eingefärbt und mit einem Haken in der Spalte „OK“ versehen. 11 Proteine wurden nach den festgelegten Kriterien ($Z \geq 98$ und $Z < 99$) nicht sicher identifiziert. Jedoch wurden 9 von diesen 11 Proteinen trotzdem korrekt an Rang 1 gelistet und sind deshalb ebenfalls mit einem Haken versehen. 33 Proteine wurden nicht identifiziert. Mit Hilfe des beschriebenen „*Scoring*“-Algorithmus wurden somit 54% aller Proteine mit einer 100 %igen Trefferquote, d.h. bei einem Z-Faktor von $Z \geq 99$ wurde auch immer das richtige Protein ermittelt, identifiziert. Das wichtigste Ergebnis dieser Studie war, dass der verwendete „*Scoring*“-Algorithmus in keinem Fall ein falsch positives Ergebnis lieferte. Die falsch negativen Ergebnisse konnten in allen Fällen auf eine zu geringe Qualität der experimentell gewonnen Daten zurückgeführt werden.

Tabelle 11 **Ergebnis der Identifizierung von 96 gentechnisch hergestellten menschlichen Proteinen in der gesamten „*NCBI* non-redundant“ Datenbank mit 686213 eingetragen Proteinen.**

Pos	Zugriffs-Nr.	Protein	OK	Z	Hits	σ [ppm]	SC [%]
A1	O95144	PARANEOPLASTIC NEURONAL ANTIGEN MA1	✓	99.5	13	10.8	58.9
C1	P26641	ELONGATION FACTOR 1-GAMMA (EF-1-GAMMA)	✓	99.3	14	6.7	25.4
E1	P04720	EF11_HUMAN ELONGATION FACTOR 1-ALPHA 1 (EF-1-ALPHA-1) (ELONGATION FACTOR TU) (EF-TU)	✓	99.5	11	6.1	49.4
G1	P18124	60S RIBOSOMAL PROTEIN L7	✓	99.3	11	6	36.7
I1	P25786	PROTEASOME COMPONENT C2 (EC 3.4.99.46) (MACROPAIN SUBUNIT C2)(PROTEASOME NU CHAIN) (MULTICATALYTIC ENDOPEPTIDASE COMPLEX SUBUNITC2) (30 KDA PROSOMAL PROTEIN) (PROS-30)	✓	99.2	8	5.6	57.4
K1	Q9ULD4	KIAA1286 PROTEIN (FRAGMENT)		92.3	6	16.5	29.9

Pos	Zugriffs-Nr.	Protein	OK	Z	Hits	σ [ppm]	SC [%]
M1	Q13509	TUBULIN BETA-3 CHAIN	✓	99.7	16	9.9	64.7
O1	Q03393	6-PYRUVOYL TETRAHYDROBIOPTERIN SYNTHASE (EC 4.6.1.10) (PTPS)(PTP SYNTHASE)	✓	99.8	18	11.7	89
A3	P22392	NUCLEOSIDE DIPHOSPHATE KINASE B (EC 2.7.4.6) (NDK B) (NDP KINASE B)(NM23-H2)	✓	99.5	7	3.1	59.9
C3	O14579	COATOMER EPSILON SUBUNIT (EPSILON-COAT PROTEIN) (EPSILON-COP)	✓	99.7	13	4.3	50.5
E3	O08765	GANGLIOSIDE EXPRESSION FACTOR 2 (GEF-2)	✓	99.2	8	7.4	69.8
G3	O75085	BAI1-ASSOCIATED PROTEIN 1 [HOMO SAPIENS]	✓	98.9	11	4.8	18.9
I3	P30533	ALPHA-2-MACROGLOBULIN RECEPTOR-ASSOCIATED PROTEIN PRECURSOR(ALPHA-2-MRAP) (LOW DENSITY LIPOPROTEIN RECEPTOR-RELATED PROTEIN-ASSOCIATED PROTEIN 1) (RAP)		94.5	6	15.5	39
K3	Q14240	EUKARYOTIC INITIATION FACTOR 4A-II (EIF-4A-II) (EIF4A-II)	✓	99.1	12	7.3	28
M3	P21333	ENDOTHELIAL ACTIN-BINDING PROTEIN (ABP-280) (NON-MUSCLE FILAMIN)(FILAMIN 1)		96.4	6	6.3	24.8
O3	P12277	CREATINE KINASE, B CHAIN (EC 2.7.3.2) (B-CK)	✓	99.9	17	2.6	58.3
A5	Q99628	SIAH BINDING PROTEIN 1 (FRAGMENT)	✓	99.6	10	3.2	38.4
C5	O15372	EIF3-P40	✓	99	10	9.4	46.3
E5	P32889	ADP-RIBOSYLATION FACTOR 1	✓	99.3	5	1.7	50.3
G5	P16152	CARBONYL REDUCTASE [NADPH] 1 (EC 1.1.1.184)	✓	99.5	11	6.5	58.7
I5	P35221	ALPHA-1 CATENIN (CADHERIN-ASSOCIATED PROTEIN) (ALPHA E-CATENIN)		97.8	8	9.6	34.2

Pos	Zugriffs-Nr.	Protein	OK	Z	Hits	σ [ppm]	SC [%]
K5	Q09028	CHROMATIN ASSEMBLY FACTOR 1 P48 SUBUNIT (CAF-1 P48 SUBUNIT)(RETINOBLASTOMA BINDING PROTEIN P48)		97.3	5	5.2	38
M5	P12277	CREATINE KINASE, B CHAIN (EC 2.7.3.2) (B-CK)	✓	99.8	18	10.3	85.8
O5	Q13867	BLEOMYCIN HYDROLASE (EC 3.4.22.-) (BLM HYDROLASE)	✓	99.9	13	2.1	49
A7	P00938	TRIOSEPHOSPHATE ISOMERASE (EC 5.3.1.1) (TIM)	✓	99	9	9.8	59
C7	P11518	60S RIBOSOMAL PROTEIN L7A (PLA-X POLYPEPTIDE) (SURF-3) [HOMO SAPIENS]	✓	98.6	6	3.8	37.6
E7	P30086	PHOSPHATIDYLETHANOLAMINE-BINDING PROTEIN (NEUROPOLYPEPTIDE H3)	✓	99.8	12	3.5	80.1
G7	P02571	ACTIN, CYTOPLASMIC 2 (GAMMA-ACTIN)	✓	99.5	12	3.4	24.8
I7	P49815	TUBERIN (TUBEROUS SCLEROSIS 2 PROTEIN)		98	14	11.7	14.6
K7	P04406	GLYCERALDEHYDE 3-PHOSPHATE DEHYDROGENASE, LIVER (EC 1.2.1.12)	✓	99.9	23	5.6	57.2
M7	P18615	RD PROTEIN		97.1	6	7.8	37.7
O7	O75769	HYPOTHETICAL 40.5 KDA PROTEIN		97.2	7	8.7	32.3
A9	O95834	ECHINODERM MICROTUBULE-ASSOCIATED PROTEIN-LIKE EMAP2		92.3	5	13.2	34.5
C9	Q01082	SPECTRIN BETA CHAIN, BRAIN (SPECTRIN, NON-ERYTHROID BETA CHAIN)(FODRIN BETA CHAIN) (SPTBN1)		98.6	13	8.3	17.9
E9	P02571	ACTIN, CYTOPLASMIC 2 (GAMMA-ACTIN)	✓	99.3	9	8.9	83.2
G9	Q9NTM3	BA124N14.1 (VIMENTIN)	✓	99.6	19	12.5	40.4
I9	P12277	CREATINE KINASE, B CHAIN (EC 2.7.3.2) (B-CK)		95	5	5.7	22.8

Pos	Zugriffs-Nr.	Protein	OK	Z	Hits	σ [ppm]	SC [%]
K9	Q9Y462	ZINC FINGER PROTEIN, ZNF6 LIKE		98.3	12	10.9	22.3
M9	Q99661	MITOTIC CENTROMERE-ASSOCIATED KINESIN	✓	99	13	11.8	35.9
O9	Q9Y6E9	SIRTUIN TYPE 2		97.1	7	9.6	33.5
A11	P02096	HEMOGLOBIN GAMMA-A AND GAMMA-G CHAINS		93.5	10	10.8	48.9
C11	P06746	DNA POLYMERASE BETA (EC 2.7.7.7)	✓	99.9	18	5	52.4
E11	Q9UIC3	NCK-2	✓	98.3	7	4.6	26.8
G11	O00318	PUTATIVE PROTEIN		93.5	6	17.9	38.6
I11	P30533	ALPHA-2-MACROGLOBULIN RECEPTOR-ASSOCIATED PROTEIN PRECURSOR(ALPHA-2-MRAP) (LOW DENSITY LIPOPROTEIN RECEPTOR-RELATED PROTEIN-ASSOCIATED PROTEIN 1) (RAP)	✓	99.4	12	5.9	32.8
K11	O14771	PUTATIVE TRANSCRIPTION FACTOR CR53 (FRAGMENT)		95.6	5	19.9	89.6
M11	P36578	60S RIBOSOMAL PROTEIN L1 (L4)	✓	99	6	3.7	53.9
O11	P27797	CALRETICULIN PRECURSOR (CRP55) (CALREGULIN) (HACBP) (ERP60) (52 KD RIBONUCLEOPROTEIN AUTOANTIGEN RO/SS-A)	✓	99.8	18	6.4	53.7
A13	P12750	40S RIBOSOMAL PROTEIN S4, X ISOFORM		90.9	6	14	21.5
C13	P07942	LAMININ BETA-1 CHAIN PRECURSOR (LAMININ B1 CHAIN)	✓	99.7	23	5.6	16.3
E13	P53041	SERINE/THREONINE PROTEIN PHOSPHATASE 5 (EC 3.1.3.16) [HOMO SAPIENS]	✓	98.7	9	7.4	36.3
G13	P78406	MRNA-ASSOCIATED PROTEIN MRNP 41 (RAE1 PROTEIN HOMOLOG)	✓	99.8	10	1.6	45.4

Pos	Zugriffs-Nr.	Protein	OK	Z	Hits	σ [ppm]	SC [%]
I13	P09651	HETEROGENEOUS NUCLEAR RIBONUCLEOPROTEIN A1	✓	99.7	12	4.3	43.4
K13	P49593	PUTATIVE PROTEIN PHOSPHATASE 2C (EC 3.1.3.16) (PP2C)	✓	99.4	9	2.8	26.9
M13	P22392	NUCLEOSIDE DIPHOSPHATE KINASE B (EC 2.7.4.6) (NDK B) (NDP KINASE B)(NM23-H2)	✓	99.6	10	7.7	93.4
O13	P31321	CAMP-DEPENDENT PROTEIN KINASE TYPE I-BETA REGULATORY CHAIN		84	7	19.1	12.2
A15	P02570	ACTIN, CYTOPLASMIC 1 (BETA-ACTIN).	✓	99.7	12	3	38.1
C15	P02768	SERUM ALBUMIN PRECURSOR	✓	99.6	9	3.5	59.5
E15	Q9UIC2	RNB6		97	5	3.4	23.2
G15	Q9ULD4	KIAA1286 PROTEIN (FRAGMENT)		93.5	5	12.8	39.4
I15	Q9UJU6	SRC HOMOLOGY 3 DOMAIN-CONTAINING PROTEIN HIP-55		97.8	8	9.6	34.2
K15	O75822	EUKARYOTIC TRANSLATION INITIATION FACTOR 3 SUBUNIT 1 (EIF-3 ALPHA)(EIF3 P35)	✓	99.8	13	4.6	55.4
M15	P49593	PUTATIVE PROTEIN PHOSPHATASE 2C (EC 3.1.3.16) (PP2C)	✓	100	24	2	48.2
O15	P12324	TROPOMYOSIN, CYTOSKELETAL TYPE (TROPOMYOSIN 3, CYTOSKELETAL) (TM30-NM)	✓	99.3	12	10.8	53.6
A17	P78549	HNTH1 (ENDONUCLEASE III HOMOLOG)		97.3	5	11.1	81.8
C17	P12277	CREATINE KINASE, B CHAIN (EC 2.7.3.2) (B-CK)	✓	99.9	17	5.6	84
E17	P04720	ELONGATION FACTOR 1-ALPHA 1 (EF-1-ALPHA-1) (ELONGATION FACTOR TU) (EF-TU)	✓	99.8	18	5.1	49.1
G17	Q9UG41	HYPOTHETICAL 65.1 KDA PROTEIN (FRAGMENT)		86.4	8	17.7	10.2

Pos	Zugriffs-Nr.	Protein	OK	Z	Hits	σ [ppm]	SC [%]
I17	P12277	CREATINE KINASE, B CHAIN (EC 2.7.3.2) (B-CK) [HOMO SAPIENS]	✓	98.3	5	1.8	21.1
K17	Q92667	A KINASE ANCHOR PROTEIN		91	7	24.3	27.5
M17	P26641	ELONGATION FACTOR 1-GAMMA (EF-1-GAMMA).	✓	99.6	18	11.5	43
O17	Q13885	BETA TUBULIN	✓	99.7	17	8.8	59.1
A19	Q13541	4E-BINDING PROTEIN 1		94.3	5	10	35.2
C19	P11142	HEAT SHOCK COGNATE 71 KDA PROTEIN	✓	99.8	23	9.7	58.2
E19	Q13838	PROBABLE ATP-DEPENDENT RNA HELICASE P47	✓	99.8	17	6.6	68.2
G19	P13639	ELONGATION FACTOR 2 (EF-2)		97.9	5	2.5	23.6
I19	P07437	TUBULIN BETA-1 CHAIN	✓	99.5	19	16	43
K19	P50502	HSC70-INTERACTING PROTEIN (HIP) (PUTATIVE TUMOR SUPPRESSOR ST13) (PROGESTERONE RECEPTOR-ASSOCIATED P48 PROTEIN)		96.8	5	12.8	78.8
M19	Q9Y4W6	PARAPLEGIN-LIKE PROTEIN		95.6	5	5.5	24.8
O19	Q13765	NASCENT POLYPEPTIDE ASSOCIATED COMPLEX ALPHA SUBUNIT		91.3	6	9.4	15
A21	P05092	PEPTIDYL-PROLYL CIS-TRANS ISOMERASE A (EC 5218)	✓	99	10	12.7	61.6
C21	Q02878	60S RIBOSOMAL PROTEIN L6 (TAX-RESPONSIVE ENHANCER ELEMENT BINDING PROTEIN 107) (TAXREB107) (NEOPLASM-RELATED PROTEIN C140) [HOMO SAPIENS]	✓	98.8	10	9.6	39.6
E21	Q09028	CHROMATIN ASSEMBLY FACTOR 1 P48 SUBUNIT (CAF-1 P48 SUBUNIT)(RETINOBLASTOMA BINDING PROTEIN P48) (RETINOBLASTOMA-BINDING PROTEIN4) (MSI1 PROTEIN HOMOLOG)		84.7	6	12.8	11.6

Pos	Zugriffs-Nr.	Protein	OK	Z	Hits	σ [ppm]	SC [%]
G21	O60833	INADL, C-TERM VARIANT2	✓	99.1	20	20	28
I21	Q12874	SPLICESOME-ASSOCIATED PROTEIN SAP 61	✓	99.3	17	24.9	59.9
K21	Q12874	SPLICESOME-ASSOCIATED PROTEIN SAP 61		96.9	9	7.9	15.7
M21	O75939	45 KDA SPLICING FACTOR		93.6	5	4.9	15.2
O21	Q9UMP5	AP4 PROTEIN		93.2	5	7.4	21.5
A23	P14373	ZINC-FINGER PROTEIN RFP (RET FINGER PROTEIN) [HOMO SAPIENS]	✓	98.9	7	5.1	47.3
C23	O14775	GUANINE NUCLEOTIDE-BINDING PROTEIN BETA SUBUNIT 5 (TRANSDUCIN BETACHAIN 5)	✓	99	8	4.2	31.4
E23	O75769	HYPOTHETICAL 40.5 KDA PROTEIN		70.1	6	25.1	11.7
G23	O00318	PUTATIVE PROTEIN [HOMO SAPIENS]	✓	98.7	6	3.8	39.2
I23	Q16643	DREBRIN E		97	11	17.3	23.6
K23	P13639	ELONGATION FACTOR 2 (EF-2)		97.5	6	5.2	28.8
M23	Q13885	BETA TUBULIN	✓	99	15	12.8	29.3
O23	P04406	GLYCERALDEHYDE 3-PHOSPHATE DEHYDROGENASE, LIVER (EC 1.2.1.12)	✓	99.2	12	9.2	40.1

4.3 Software Paket MS-Proteomics

4.3.1 Grundlegende Funktion und Eigenschaften

MS-Proteomics ist eine komplette „*Client/Server*"-Anwendung [78-83], die es dem Benutzer ermöglicht, mehrere hundert Proteinmassenspektren in kurzer Zeit zu analysieren. Mit Hilfe des „*Clients*" wählt der Benutzer über verschiedene Auswahlmasken (Abbildung 33) alle relevanten Parameter (z.B. Datenbank in der die Suche erfolgen soll , Spezies , Modifikationen u.a.) aus und startet die Suche.

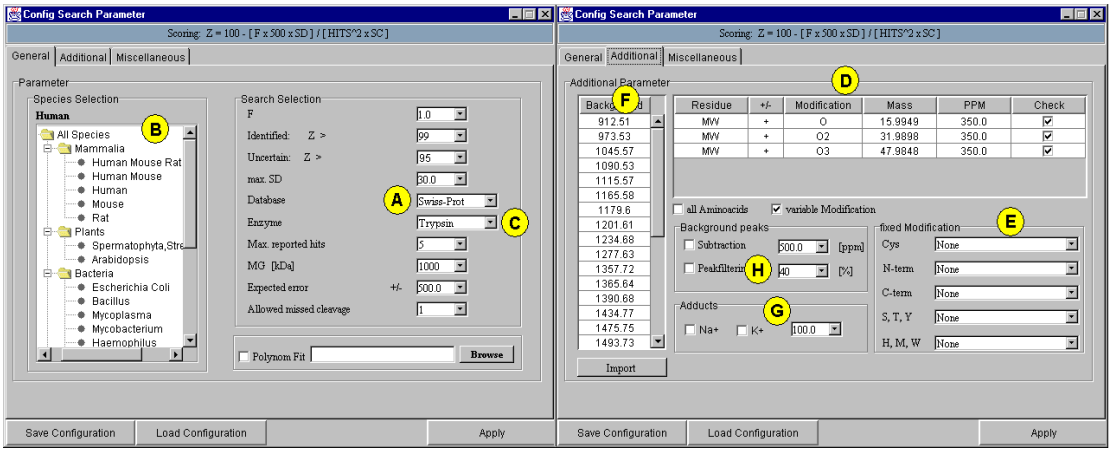

Abbildung 33 Screenshot von zwei Auswahlmasken der „*Client*"-Anwendung des Softwarepakets MS-Proteomics.
Im folgenden werden beispielhaft einige Auswahlmöglichkeiten beschrieben. Der Benutzer kann zwischen verschiedenen Sequenzdatenbanken ausgewählten (A). Die Suche kann auf eine bestimmte Spezies innerhalb der unter A ausgewählten Datenbank beschränkt werden (B). Es stehen mehrere Spaltenzyme zur Auswahl (C). Es besteht die Möglichkeit bei der Suche erwartete variable (D) und/oder permanente (E) Modifikationen festzulegen. Es können bestimmte Peptidmassen von der Suche ausgeschlossen werden (F). Weiterhin ist es möglich Na^+ und/oder K^+ Kontaminationen (G) und Massen, die sich zu einem gewissen Prozentsatz in allen zur Suche verwendeten Peaklisten befinden, herauszufiltern (H).

Der „*Server*" empfängt die von dem „*Client*" gesendeten Daten über ein „Servlet" [84,85], generiert aus den erhaltenen Daten eine Datenbankabfrage,

führt u.a. die unter 4.1 beschriebenen Schritte 1-5 aus, ermittelt den Scoring-Faktor (siehe 4.2) und sendet die ermittelten Resultate zu dem anfragenden „Client“ zurück. Der „Server“ ist in der Lage, Anfragen von mehreren Clients gleichzeitig zu bearbeiteten. Das Programm bietet dem Benutzer verschiedene Filtermöglichkeiten, um z.B. bei der Probenpräparation entstandene Verunreinigungen zu erkennen und von der Suche auszuschließen:

- In eine Liste können bestimmte Peptidmassen (z.B. von tryptischen Keratinpeptiden) eingetragen werden, die unter Berücksichtigung eines vom Benutzer festgelegten Fehlerintervalls aus allen Peaklisten herausgefiltert werden (Abbildung 33-F).
- Peptide, die beim MALDI-Prozeß statt des erwarteten Protons ein Na^+ - und/oder ein K^+-Ion aufgenommen haben, können ebenfalls unter Berücksichtigung eines vom Benutzer festgelegten Fehlerintervalls aus allen Peaklisten herausgefiltert werden (Abbildung 33-G).
- Eine weiterer optionaler Filter ermöglicht es dem Benutzer, Massen zu selektieren, die sich zu einem gewissen Prozentsatz in allen für eine Suche verwendeten Peaklisten befinden (Abbildung 33-H) und daher als nicht oder nur wenig spezifische Ausgangsdaten für die anschließende Datenbanksuche erachtet werden können.

Die Festlegung möglicher variablen Modifikationen wird wie folgt vorgenommen:

- Alle Aminosäurereste sind frei selektierbar.
- Die Festlegung der jeweiligen Modifikation, erfolgt entweder durch Eingabe der chemischen Summenformel, das Programm berechnet die entsprechende Masse selbständig oder die mit der Änderung verbundene Massendifferenz wird direkt eingegeben.
- Für jeden gewählten Aminosäurerest kann die entsprechende Masse addiert oder subtrahiert werden.

- Es sind mehrere Modifikationen gleichzeitig wählbar.
- Es sind mehrere Aminosäurereste gleichzeitig wählbar.
- Für jede gewählte Modifikation sind unterschiedliche Fehlerintervalle einstellbar.

Gefundene Modifikationen, werden nur dann als zusätzliche Treffer dem jeweiligen Protein zugesprochen, wenn die Peptidmasse von der die Modifikation abgeleitet wurde, auch als Treffer (siehe 4.1, Schritt 5) gewertet wurde.

Die Hauptaufgabe des „*Clients*“ besteht darin, die vom „*Server*“ übermittelten Ergebnisse möglichst übersichtlich darzustellen, so dass der Benutzer deren Interpretation ohne zusätzlichen Aufwand vornehmen kann. Dazu wurde eine graphische Oberfläche entwickelt, die alle relevanten Ergebnisse nebeneinander in einzelnen, in ihrer Größe frei skalierbaren, Fenstern darstellt (Abbildung 34 A-G, Abbildung 35 I) und dem Benutzer interaktiv die Möglichkeit bietet, relevante Zusammenhänge schnell zu analysieren. In der gesamten Anwendung wurde analog zu den unter 4.2 beschriebenen Identifizierungskriterien eine einheitliche und eindeutige Farbkodierung gewählt.

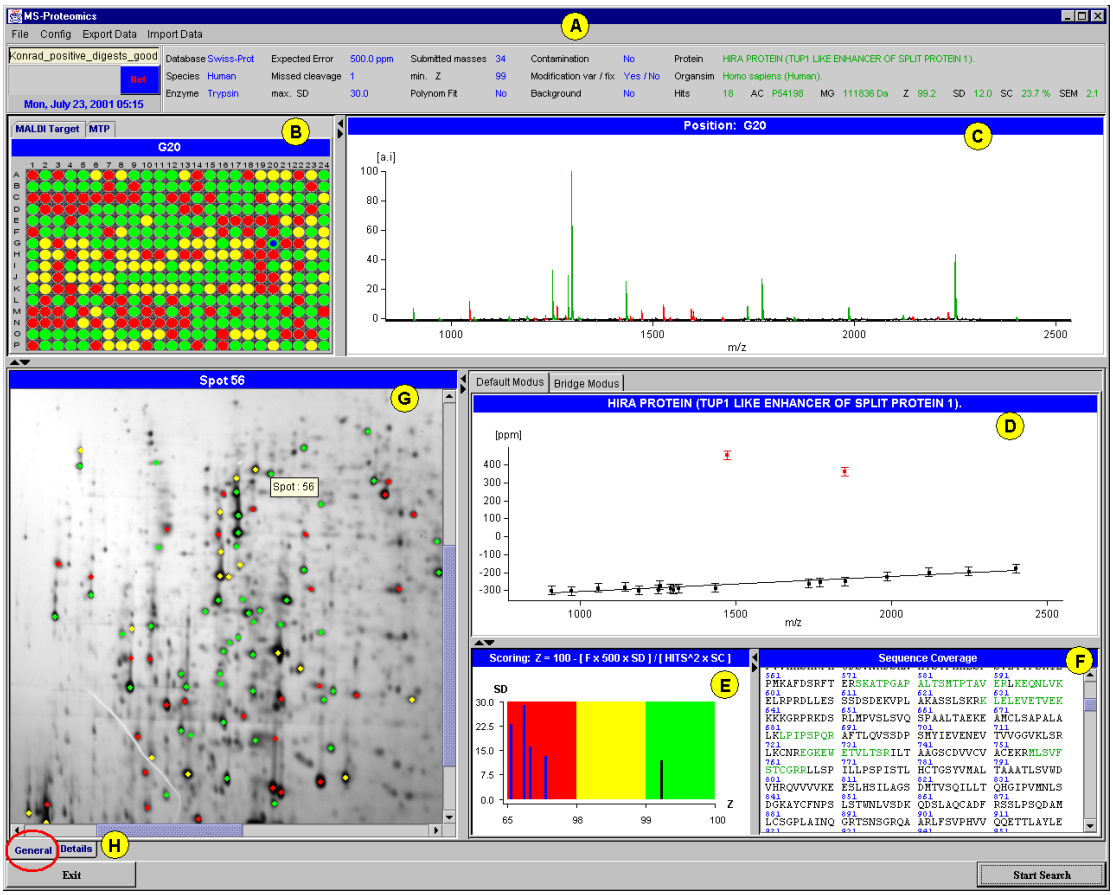

Abbildung 34 Screenshot I von der Oberfläche des Softwarepakets MS-Proteomics. Erklärung siehe Text und Tabelle 12 .

Über die in Abbildung 34 markierte Tabulatortaste H besteht die Möglichkeit zwischen Ansicht D-G und Ansicht I (Abbildung 35) umzuschalten.

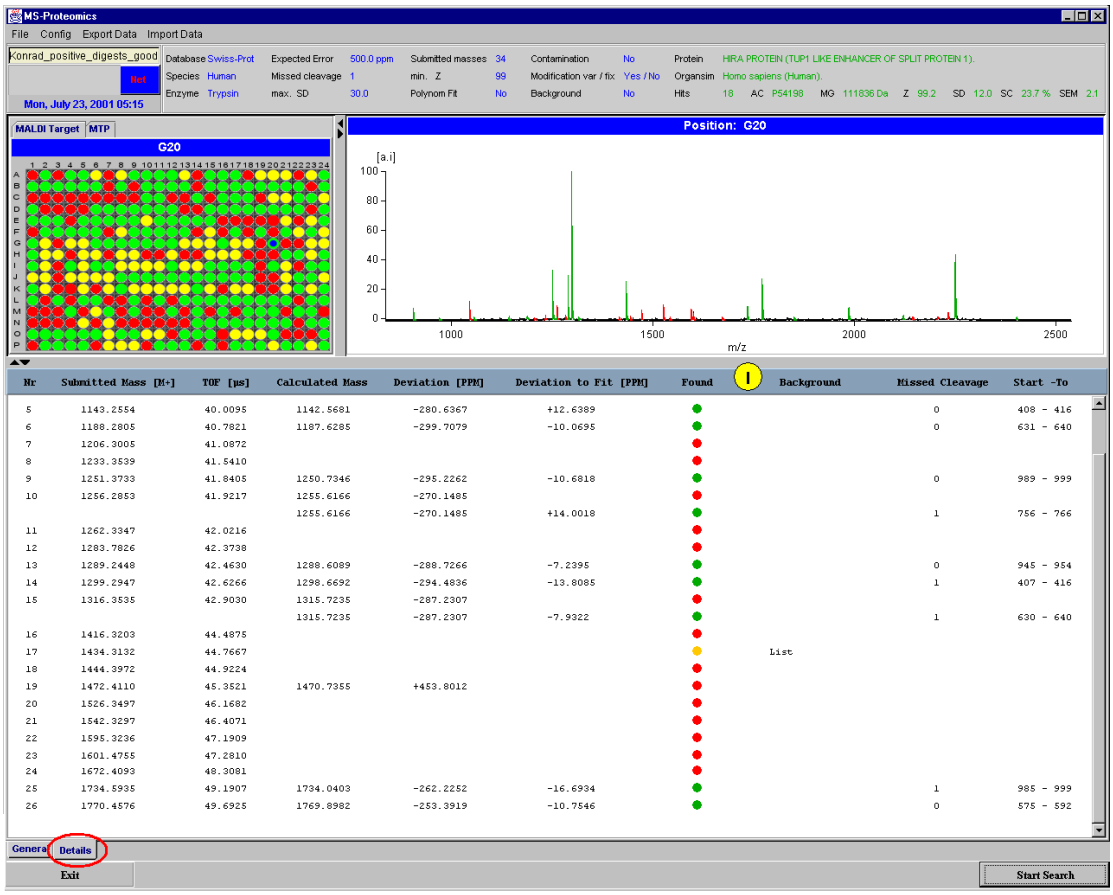

Abbildung 35 Screenshot II von der Oberfläche des Softwarepakets MS-Proteomics. Erklärung siehe Text und Tabelle 12.

In Tabelle 12 werden die grundlegenden Funktionen und Eigenschaften der in Abbildung 34 markierten Fenster A-G und des in Abbildung 35 markierten Fensters I beschrieben.

Tabelle 12 Funktion und Eigenschaften der Fenster A-G in Abbildung 34 bzw. von Fenster I in Abbildung 35.

Fenster	Funktion/Eigenschaften
A	Auflistung vom Benutzer gewählten Suchparameter.
B	Schematische Darstellung der 384er Mikrotiterplatte bzw. des 384er MALDI-Probenträgers, mit den jeweiligen Proben auf ihren entsprechenden Positionen. Beide Ansichten sind über Tabulatortasten anwählbar. Die Position können unterschiedlich eingefärbt sein: Wenn keine Probe aufgetragen wurde, ist die Position grau gefärbt (nicht gezeigt). Die Farbe grün steht für eindeutig identifiziert ($Z \geq 99$), gelb für wahrscheinlicher Kandidat, für dessen

Fenster	Funktion/Eigenschaften
$B_{Fortsetzung}$	Identifikation aber zusätzliche Informationen benötigt werden ($Z \geq 98$ und $Z < 99$) und rot für nicht identifiziert ($Z < 98$).
C	Graphische farbige Darstellung der Massenspektren. Grüne Farbe: Peak wurde beim Peakpicking als Peptidpeak erkannt und dem entsprechenden Protein als Treffer zuerkannt. Rote Farbe: Peak ist vermeintlicher Peptidpeak ergab aber keinen Treffer. Blaue Farbe: Peaks, die einer variablen Modifikation entsprechen. Mögliche Modifikationen wurden vom Benutzer vor dem Start der Suche über eine Auswahlmaske (Abbildung 34) festgelegt. Orange Farbe: Peaks symbolisieren Verunreinigungen, die über eine Auswahlmaske (Abbildung 33-F,G,H) spezifiziert wurden und automatisch anschließend vom Programm herausgefiltert werden.
D	Graphische Darstellung des Verhältnisses von der ermittelten Abweichung in ppm versus m/z (siehe 4.1, Schritt 3 und 4).
E	„*Scoring*“ Fenster: Graphische Darstellung des Verhältnisses von der ermittelten Standardabweichung (siehe 4.1, Schritt 4) zu dem „*Scoring“*-Faktor Z der gefunden Proteine (siehe 4.2). Die x-Achse ist in drei farblich unterschiedlich markierte Bereiche unterteilt: Rote Fläche: zwischen Z_0 = kleinster Z-Faktor der gefunden Proteine bis $Z_n < 98$. Gelbe Fläche: zwischen $Z_0 \geq 98$ und $Z_n < 99$. Grüne Fläche: zwischen $Z_0 \geq 99$ und $Z_n \leq 100$.
F	Darstellung der Proteinsequenz. Peptidesequenzen von Treffern sind grün markiert.
G	Optionale graphische Darstellung des Gels. Die einzelnen Gelpositionen sind den in Fenster B verwendeten Probenpositionen zugeordnet und analog dazu auch eingefärbt. In der aktuellen Suche nicht verwendete Gelpositionen sind rosa markiert.
I	Tabellarische Darstellung der Ergebnisse für das aktuell ausgewählte Protein. 1. Spalte: Nummer des Peaks. 2. Spalte: Gemessene Masse (protoniert, monoisotopisch). 3. Spalte: Flugzeit in µs.

Fenster	Funktion/Eigenschaften
$I_{Fortsetzung}$	4. Spalte: Gefundene Masse (nicht protoniert, monoisotopisch). 5. Spalte: Abweichung der gemessenen Masse von der gefundenen, berechneten Masse in ppm (analog zu 4.1, Schritt 2). 6. Spalte: Abweichung der gemessenen Masse von der korrigierten Masse in ppm (analog zu 4.1, Schritt 3). 7. Spalte: Farbige Markierung der Peakzuordnung, Farbkodierung analog zu Fenster C 8. Spalte: Hintergrundsignale bzw. variable Modifikationen . 9. Spalte: Anzahl der vom verwendeten Spaltenzym übersehenen Spaltstellen. 10. Spalte: Start und Ende der jeweiligen Peptidsequenz in der Proteinsequenz.

Die jeweiligen Eigenschaften bzw. Funktionen der oben beschriebenen Fenster sind getrennt von deren Erscheinungsbild in unterschiedlichen Klassen [86,87] definiert. Die Modularität der einzelnen Komponenten ermöglicht dem Entwickler ohne großen Aufwand, das vorhandene Programm um neue Funktionen zu erweitern bzw. vorhandene Funktionen mit neu entwickelten graphischen Objekten zu kombinieren. Die hier beschriebene „*Client*"-Anwendung ist ein sogenanntes „*Multithreading*"-Programm [88,89]. Bei Programmstart wird jedem Fenster ein eigener „*Thread*" [90-94] zugeordnet. Jeder „*Thread*" läuft in seinem eigenen Kontext ab, d.h. einzelne Programmteile können unabhängig voneinander agieren. Der Benutzer hat damit z.B. die Möglichkeit eine Suche mit mehreren hundert Proben zu starten und während die Suche noch andauert, bereits erhaltene Ergebnisse zu analysieren oder ein Gelbild zu laden ohne das die Anwendung blockiert wird. Die einzelnen oben beschriebenen Fenster stehen über einige wenige sogenannte statische Variabeln miteinander in Verbindung. Durch Klicken mit der rechten Maustaste auf eine bestimmte Position der Mikrotiterplatte bzw. des Probenträgers (Abbildung 34-B) erscheint ein Popup-Menü (nicht gezeigt), in dem die Namen aller für diese Probenposition gefunden Proteine, nach ihrem „*Scoring*"-Faktor geordnet, aufgelistet sind. Durch Auswählen eines Listeneintrags werden alle anderen Fenster auf den neusten Stand

gebracht. Im Spektrumfenster (Abbildung 34-C) wird z.B. das entsprechende Massenspektrum für das aus der Liste gewählte Protein geladen und dargestellt. Durch Ziehen mit der Maus besteht die Möglichkeit Teilbereiche des Spektrums oder das Isotopenmuster eines Peaks vergrößert darzustellen. Falls die Proben von einem Gel stammen, wird der entsprechende Gelspot (Abbildung 34-G) markiert und umgekehrt durch Klicken auf einen bestimmten Gelspot werden die anderen Fenster erneut aktualisiert. Die Interaktivität des Programms ist auch auf der Peptidebene gegeben, z.B. durch Positionierung des Mauszeigers auf einen einzelnen Punkt im Plotfenster (Abbildung 34-D) eine Zeile im Detailfenster (Abbildung 35-I, 7.Spalte) wird sowohl die entsprechende Peptidsequenz im Sequenzfenster (Abbildung 34-F), als auch der zugehörige Peak im Spektrumfenster (Abbildung 34-C) markiert.

4.3.2 Erweiterte Funktion und Eigenschaften

Das Programm bietet dem Benutzer die Möglichkeit vom „*Server*“ gelieferte Ergebnisse in einem programmeigenen Binärformat abzuspeichern, die dann bei Bedarf, ohne erneuten Zugriff auf die Datenbank, weiterverarbeitet werden können. Um die Daten anderen Anwendungen zur Verfügung zu stellen besteht darüber hinaus die Möglichkeit die Ergebnisse im ASCII-Textformat in die Zwischenablage des Betriebssystems zu kopieren. Auf die gleiche Weise besteht die Möglichkeit auf die Aminosäuresequenz ausgewählter Proteine durch Sequenzeditorprogramme (z.B. GPMAW von Lighthouse data) zuzugreifen.

4.4 Analyse ausgewählter 2D-Gelspots

Aus einem von einer Probe menschlichen Gehirns angefertigten 2D-Gel wurden 15 Spots unterschiedlicher Größe und Färbung (nummeriert in Abbildung 36 1-15) wie unter 3.2.4 beschrieben prozessiert.

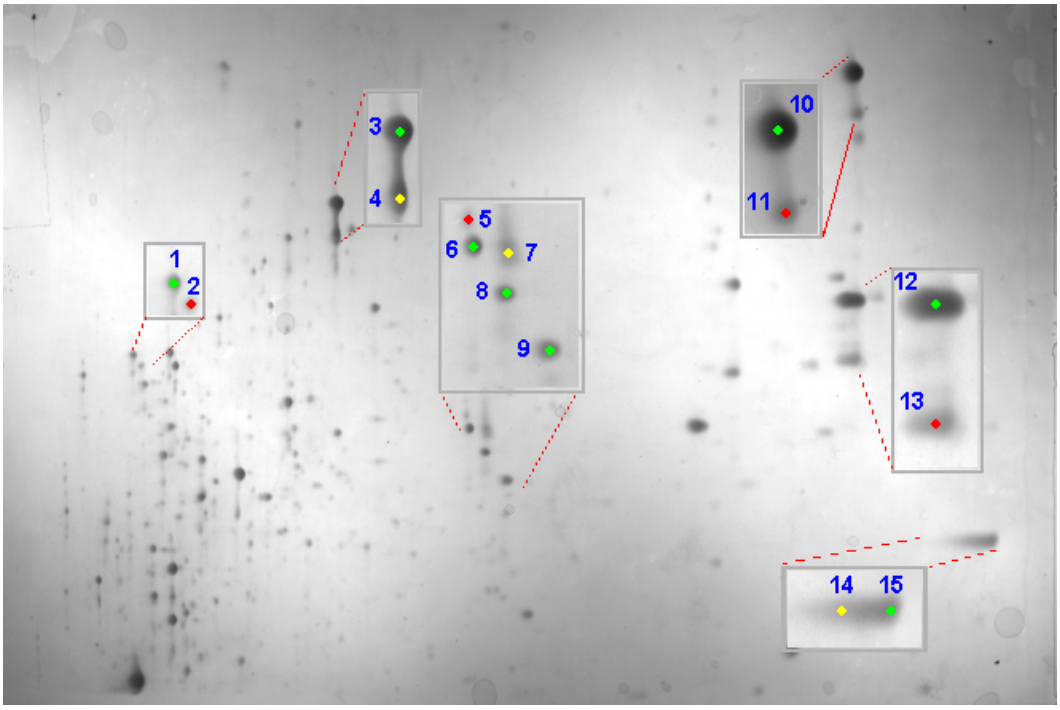

Abbildung 36 2D-Gel einer Probe aus menschlichem Gehirn.
Proben 1-15 wurden analysiert. Das Gel wurde freundlicherweise von Prof. Dr. Dr. J. Klose, Institut für Humangenetik, Forschungshaus, Charite CVK, Berlin, zur Verfügung gestellt.

Die ausgewählten Proben wurden wie unter 3.2.4 beschrieben aufgearbeitet auf einem 384er MALDI-Probenträger aufgetragen (Abbildung 37).

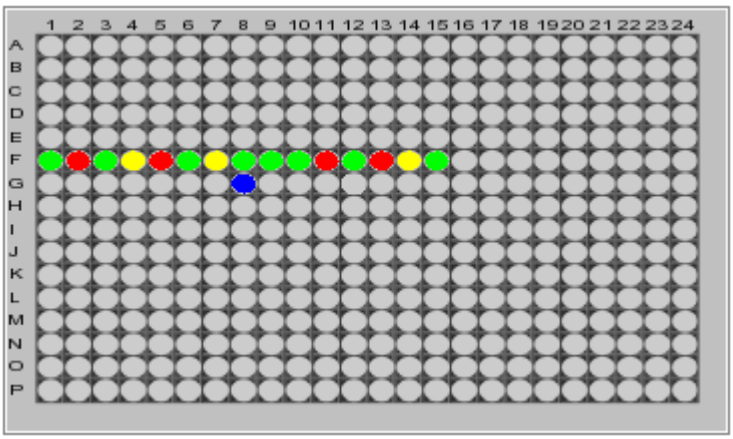

Abbildung 37 Schematische Darstellung des 384er MALDI-Probenträgers.
Der PPG-Standard wurde an Position G8 (blau markiert) aufgetragen. Die Probenpositionen (F1-15 analog zu Abbildung 36) sind entweder grün, gelb oder rot markiert. Grün symbolisiert, dass die entsprechende Probe eindeutig

identifiziert wurde ($Z \geq 99$), gelb ($Z \geq 98$ und $Z < 99$) steht für nicht sicher identifiziert und rot ($Z < 98$) für nicht identifiziert.

Die Probenpräparation wurde wie unter 3.2.2, das „*Peakpicking*“ wie unter 3.2.1 beschrieben durchgeführt. Die Massen wurden mit Hilfe einer Polynomfunktion 15^{ten} Grades aus den bestimmten Flugzeiten für welche die 16 Konstanten a_0-a_{15} mit Hilfe des PPG-Standards bestimmt wurden, berechnet (siehe 4.3). Die Datenbanksuche erfolgte unter den gleichen, oben bereits beschriebenen Bedingungen. Zusätzlich wurden mögliche Oxidationen der Aminosäuren Methionin und Tryptophan berücksichtigt. Weiterhin wurden Peptidmassen, die bei einer erlaubten Fehlertoleranz von ± 500ppm in mehr als 50% aller generierten Peaklisten und/oder in einer Liste von ausgewählten Keratinpeptidmassen vorhanden waren, vor Beginn der Suche herausgefiltert. Die Suche wurde in der Teildatenbank „Mammalia” der „NCBI non-redundant“ Datenbank durchgeführt. Die Ergebnisse sind in Tabelle 13 aufgelistet.

Tabelle 13 **Ergebnis der Identifizierung von 15 durch 2D-Gelelkrophorese getrennten Proben aus menschlichem Gehirn. Die Suche wurde in der Teildatenbank „Mammalia” der „NCBI non-redundant“ Datenbank durchgeführt.**

Spot	Pos	Zugriffs-Nr.	Protein	OK	Z	Hits	σ [ppm]	SC [%]
1	F1	O00154	CYTOSOLIC ACYL COENZYME A THIOESTER HYDROLASE (EC 3.1.2.2) **[HOMO SAPIENS]**	✓	99.3	17	17,4	45.2
2	F2	NP_060563.1	HYPOTHETICAL PROTEIN FLJ10439 [HOMO SAPIENS] GI 7022471 DBJ BAA91610.1 (AK001301) UNNAMED PROTEIN PRODUCT **[HOMO SAPIENS]**		95.5	5	3.6	15.8
3	F3	XP_009216.1	TRANSCRIPTION FACTOR **[HOMO SAPIENS]**	✓	99.2	5	1.1	27.9
4	F4	XP_002177.1	HYPOTHETICAL PROTEIN FLJ10845 **[HOMO SAPIENS]**	✓	98.2	9	6.7	23.6
5	F5	NP_064327.1	TESTIS-SPECIFIC POLY(A) POLYMERASE **[MUS MUSCULUS]**		94.7	7	11.2	21.7
6	F6	PIR I56489	NEUROPOLYPEPTIDE H3, BRAIN – HUMAN (FRAGMENT) **[HOMO SAPIENS]**	✓	99.5	8	3.7	55

7	F7	(AF263308)	RAB3 INTERACTING PROTEIN VARIANT 4 **[HOMO SAPIENS]**	✓	98	6	2.3	15.9

Spot	Pos	Zugriffs-Nr.	Protein	OK	Z	Hits	σ [ppm]	SC [%]
8	F8	EMB CAB46022.1 (AL023553) DJ347H13.1	ACONITATE HYDRATASE, MITOCHONDRIAL PRECURSOR (EC 4.2.1.3) **[HOMO SAPIENS]**	✓	99.5	22	21.9	44.5
9	F9	(AC005388)	SIMILAR TO GB U51990 HPRP18 (SPLICING FACTOR) GENE FROM **[HOMO SAPIENS**	✓	99.4	11	4.4	31.2
10	F10	P01922	HEMOGLOBIN ALPHA CHAIN. **HOMO SAPIENS (HUMAN), PAN TROGLODYTES (CHIMPANZEE)**	✓	99.7	8	3	74.5
11	F11	NP_004601.1	T-COMPLEX 10 (A MURINE TCP HOMOLOG) **[HOMO SAPIENS]**		87	5	14.6	22.4
12	F12	NP_000550.1	HEMOGLOBIN, GAMMA A **[HOMO SAPIENS]**	✓	99.2	10	13	83.7
13	F13	(AF186109)	TPM4-ALK FUSION ONCOPROTEIN TYPE 2 **[HOMO SAPIENS]**		97.1	7	10.3	36.3
14	F14	NP_035164.1	TRX DEPENDENT PEROXIDE REDUCTASE 2; **[MUS MUSCULUS]**	✓	98.8	9	10.6	55.3
15	F15	NP_006417.1	DIHYDROPYRIMIDINASE RELATED PROTEIN-4 (DRP-4) **[HOMO SAPIENS]**	✓	99.5	21	21.6	49.5

Im Folgenden werden anhand drei ausgewählter Proteinspots die übersichtliche Darstellung der Ergebnisse und die interaktiven Fähigkeiten von MS-Proteomics sowie der damit verbundene mögliche Erkenntnisgewinn demonstriert. In Abbildung 38 sind die Ergebnisse für die Analyse von Spot 1 gezeigt.

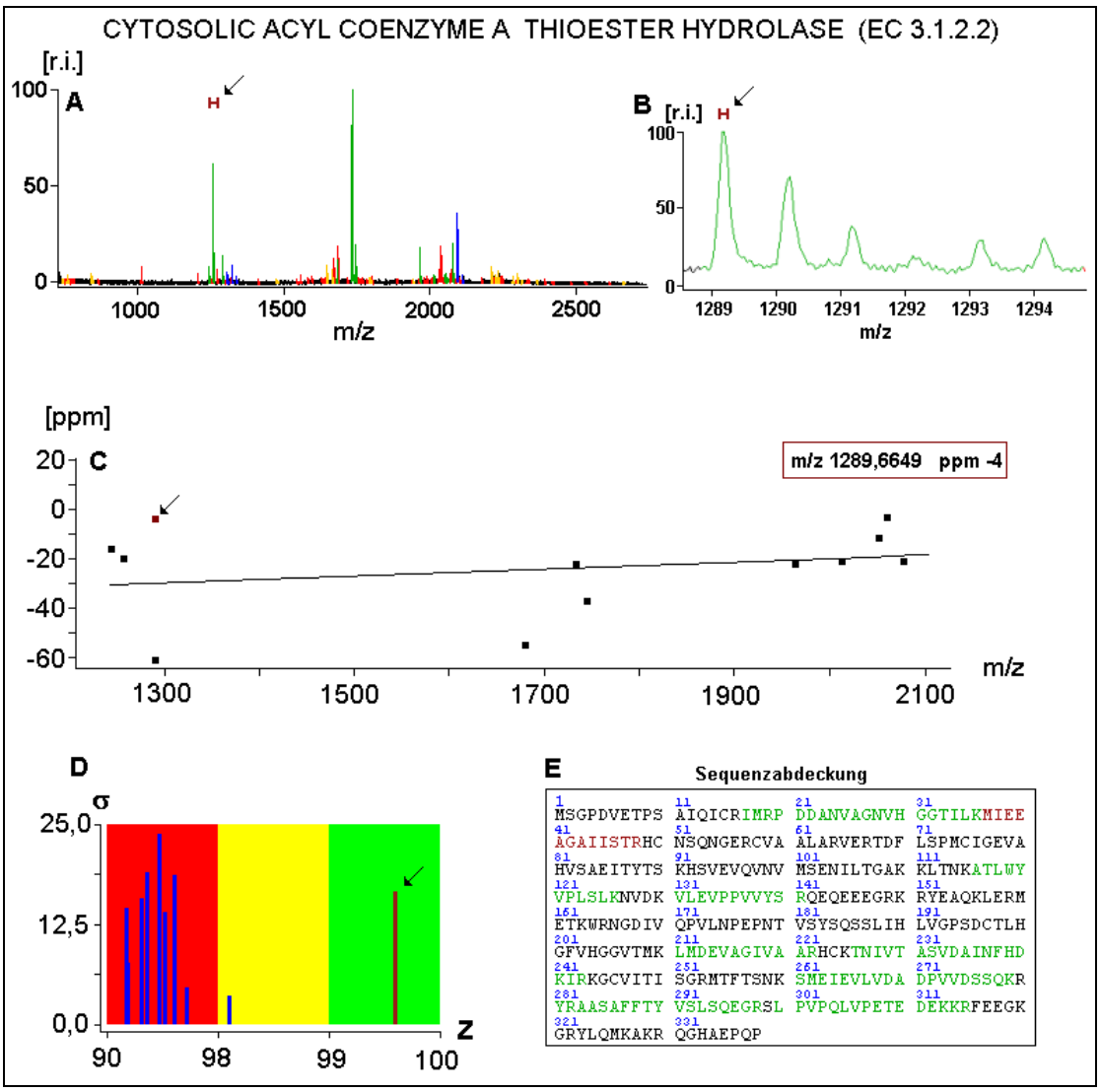

Abbildung 38 Ergebnisse für die Analyse von proteinspot 1. Erklärung siehe Text.

In dem Balkendiagramm des „*Scoring*"-Fensters (Abbildung 38-D) sind für die ersten 10 Kandidaten der Suche das jeweilige Verhältnis von der ermittelten Standardabweichung (siehe 4.1, Schritt 4) zu dem „Scoring"-Faktor Z der Proteine graphisch dargestellt. Aus dem Diagramm ist sofort zu ersehen, dass sich der das eindeutig identifizierte Protein „*Cytosolic Acyl Coenzyme A Thioester Hydrolase*" präsentierende Balken „weit" rechts im grünen Bereich des Diagramms befindet, alle anderen befinden sich im gelben bzw. roten Bereich. Wie oben bereits beschrieben, wird die Farbe grün in der gesamten Anwendung für eindeutig identifizierte Kandidaten ($Z \geq 99$) verwendet. Durch

Mausklick auf den Balken werden zusätzliche Daten für dieses Protein geladen und alle anderen Fenster aktualisiert. In Abbildung 38-A ist das zugehörige Massenspektrum dargestellt. Durch Selektion einer Peptidmasse im Plotfenster (Abbildung 38-C) wird der die Masse symbolisierende Punkt braun eingefärbt (in Abbildung 38-C mit Pfeil markiert) und gleichzeitig wird der zugehörige Massenwert (1289,6649 m/z) und die Abweichung dieser Peptidmasse von ihrer korrespondierenden gemessenen Masse (ppm -4) oben rechts im gleichen Fenster eingeblendet. Weiterhin wird im zugehörigen Spektrum der entsprechende Peak durch ein ebenfalls braun gefärbtes „H" gekennzeichnet (in Abbildung 38-A mit Pfeil markiert) und im Sequenzfenster die entsprechende Peptidsequenz braun eingefärbt (Abbildung 38-E). Durch Ziehen mit der Maus kann der im Spektrum markierte Peak so aufgelöst werden, dass die einzelnen Isotope sichtbar werden (Abbildung 38-B). In dieser Ansicht erkennt der Benutzer sofort, ob die Peakerkennungssoftware (siehe 3.2.1) wie im gezeigten Beispiel das richtige Isotop (durch braun gefärbtes „H" markiert) ausgewählt hat. Durch die übersichtliche und interaktive Darstellung ist der Benutzer in der Lage schnell Suchen seine Ergebnisse zu analysieren.

In Abbildung 39 ist der in Abbildung 38-A markierte, die Masse 1289,6649 m/z präsentierende Peak mit drei zusätzlichen Satellitenpeaks gezeigt. Die blaue Farbe signalisiert dem Benutzer, dass diese Satellitenpeaks die Modifikation einer oder mehrerer Aminosäuren eines detektierten Peptids (grün gefärbter Peak in Abbildung 39) anzeigen.

Die Art der Modifikation ist in der Spalte „Background" der Detailansicht einzusehen (analog Abbildung 35-I, für das diskutierte Beispiel nicht gezeigt).

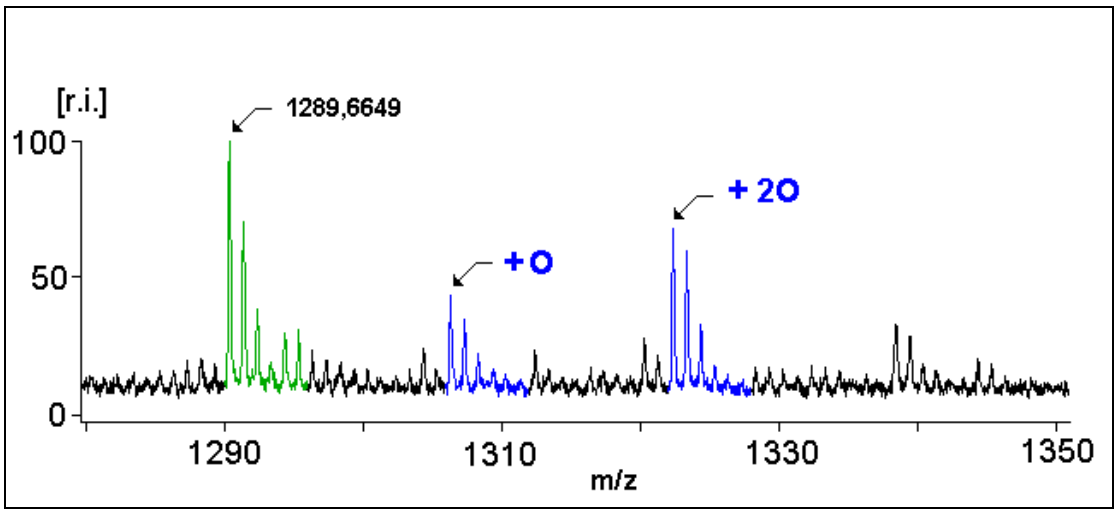

Abbildung 39 Beispiel für Kennzeichnung einer vorhanden Modifikation.
Der die gefundene Peptidmasse symbolisierende Peak ist grün gefärbt, die zugehörigen die Modifikation (im Beispiel Oxidationen) präsentierenden Peaks sind blau gefärbt

Im gezeigten Beispiel resultieren die beiden Satellitenpeaks sowohl aus einer einfachen, als auch eine zweifachen Oxidation der Aminosäure Methionin, die Bestandteil der gefundenen Masse (m/z=1289,6649) zugeordneten Peptidsequenz („**M**IEEAGAIISTR") ist (Abbildung 40).

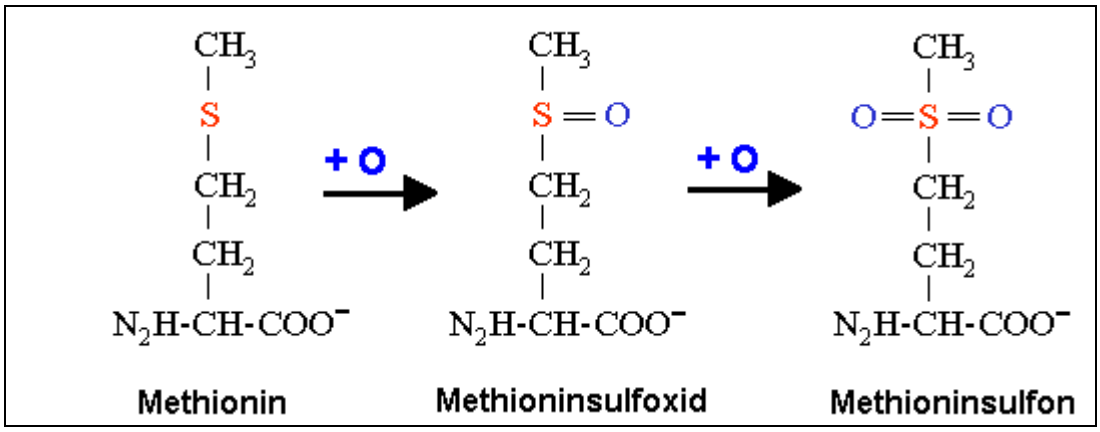

Abbildung 40 Oxidation von Methionin zu Methioninsulfoxid und Methioninsulfon.

Die bei der Suche maximal erlaubte Fehlertoleranz für eine vorhandene Methioninoxidation betrug im gezeigten Beispiel ± 50ppm. In Tabelle 14 sind für die beiden oben erwähnten Oxidationen die zugehörigen Massen und deren Abweichung von der Masse des gefundenen Peptids in ppm aufgelistet.

Tabelle 14 Angaben zur Methioninoxidation im gezeigten Beispiel

Gemessene Masse m1 [Da]	Art der Modifikation	Massenbilanz der Modifikation m2 [Da]	m3 = (m1 – m2) [Da]	Abweichung (m3 – m4) # [ppm]
1305,6202	einfach oxidiert	+ 15,9949	1289,6253	+ 30,7
1321,6635	zweifach oxidiert	+ 31,9898	1289,6737	-6,8

Masse des gefunden Peptids m4 =1289,6649 m/z (siehe oben).

Wie oben bereits erwähnt, werden gefundene Modifikationen nur dann als zusätzliche Treffer dem jeweiligen Protein zugesprochen, wenn die Peptidmasse von der die Modifikation abgeleitet wurde, auch als Treffer (siehe 4.1, Schritt 5) gewertet wurde. Dies ist insofern gerechtfertigt, weil die zusätzlichen, aufgrund der Modifikation vorhandenen Peaks die Wahrscheinlichkeit, dass es sich bei der dem nichtmodifizierten Peptid zugeordneten Masse wirklich um einen Treffer handelt, erhöhen. Enthält im Gegensatz zum gezeigten Beispiel das Peptid nicht mindestens ein Methionin- oder Tryptophan findet die Zuordnung der Satellitenpeaks nicht statt. Das heißt, die Satellitenpeaks beinhalten Informationen welche die mögliche Aminosäurezusammensetzung für das unmodifizierte Peptid einschränken. Dieser Information wird durch einen zusätzlichen Treffer Rechnung getragen. Im vorliegenden Beispiel wurden 11 Treffer ohne Berücksichtigung von variablen Modifikationen und aufgrund vorhandener Oxidationen 6 zusätzliche Treffer dem Protein zugesprochen.

Am Beispiel der Analyse von Proteinspot 15 werden die unterschiedlichen Filteroptionen des Programms MS-Proteomics veranschaulicht. Abbildung 41 zeigt das Massenspektrum der tryptisch verdauten Spaltpeptide des Proteins „Dihydropyrimidinase Related Protein-4 (DRP-4)“ in unterschiedlichen Vergrößerungen. Die mit gestrichelter Linie umrahmten Bereiche sind jeweils in der darunterliegenden Abbildungen vergrößert dargestellt. Abbildung 41-C zeigt hochaufgelöst den in Abbildung 41-B markierten Bereich zwischen m/z 2209 und 2321.

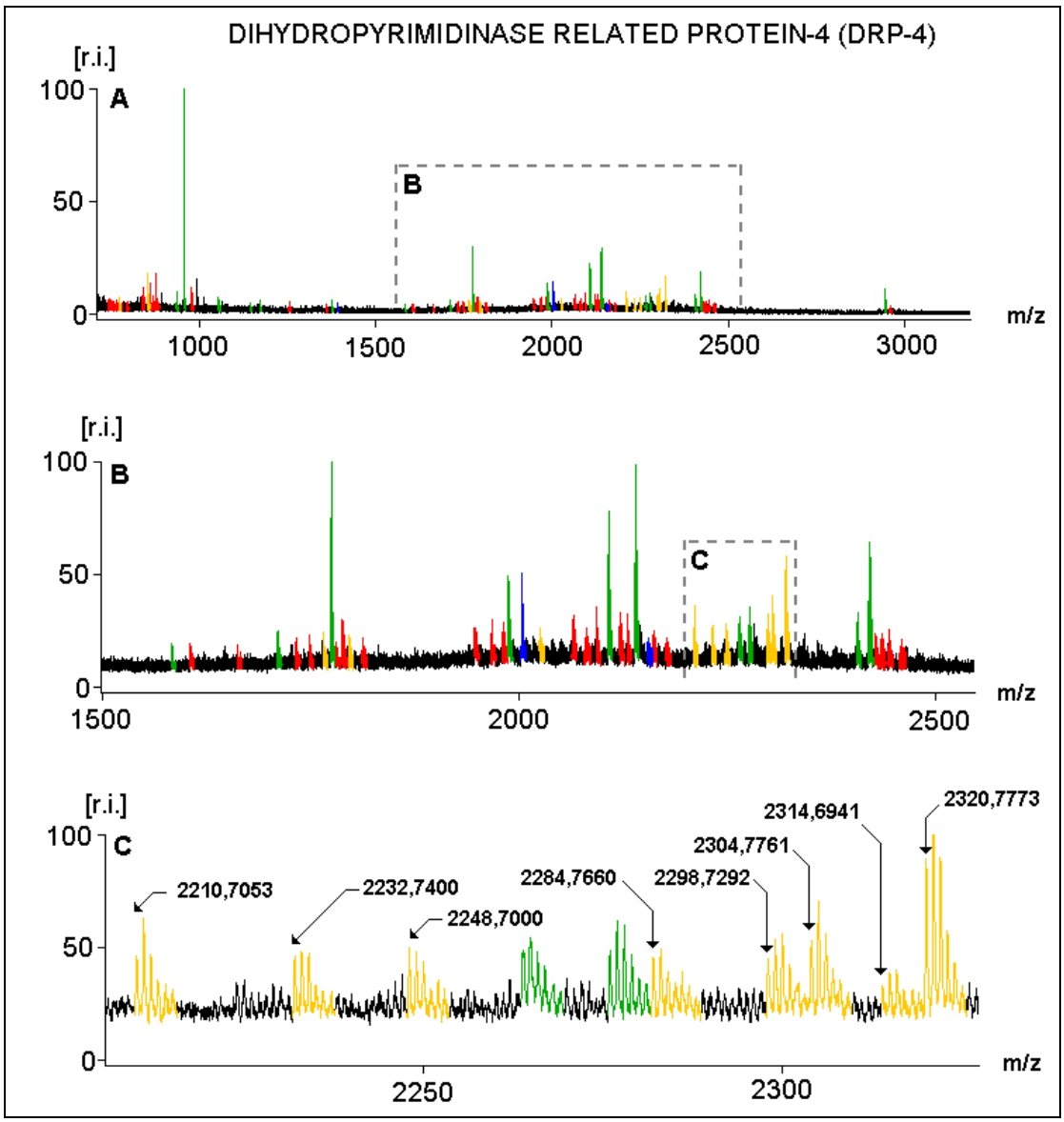

Abbildung 41 Beispiel für die Visualisierung detektierter Verunreinigungen.
Erklärung siehe Text.

Die in Abbildung 41-C 8 gelb gefärbten Peaks wurden Verunreinigungen zugeordnet. Im Gegensatz zu möglichen Modifikationen, die erst auf der Serverseite Berücksichtigung finden, werden mögliche Verunreinigungen bereits durch den „*Client*“ erkannt und die entsprechenden Massen aus der Peakliste entfernt. Die so bereinigte Peakliste wird dann an den „*Server*“ gesendet. Die Art der erkannten Verunreinigung kann vom Benutzer ebenfalls in der Detailansicht eingesehen werden (nicht gezeigt). In Tabelle 15 sind für

die 8 Peaks die entsprechenden Massen und die Art der Verunreinigung aufgelistet.

Tabelle 15 Angaben zu gefunden Verunreinigungen im gezeigten Beispiel

Gemessene Masse m1 [Da]	Keratinpetidmasse m2 [Da]	Masse in >50% aller Spektern vorhanden	Abweichung (m1 – m2) [ppm]
2210,7053	2210.1021	✓	272,9
2232,7400	2232.1022	✓	285,7
2248,7000	2248.1021	✓	265,9
2284,7660	—	✓	—
2298,7292	—	✓	—
2304,7761	2304.1722	—	262,0
2314,6941	—	✓	—
2320,7773	2320.1721	✓	260,8

Neben den oben gezeigten Filteroptionen besteht darüber hinaus die Möglichkeit Na^+- und/oder K^+-Kontaminationen aus den Peaklisten herauszufiltern. Aus der für das gezeigte Protein vorhandenen Peakliste wurden insgesamt 11 Massen herausgefiltert. Die relativ hohe Anzahl an herausgefilterten Massen zeigt, dass die beschriebenen Filteroptionen für die Identifizierung des richtigen Proteins von entscheidender Bedeutung sein können. Belässt man die Massen in der Peakliste, so besteht die Möglichkeit, dass einige oder alle dieser Massen einem anderen Protein als Treffer zugeschrieben werden und dieses somit fälschlicherweise einen höheren Z-Faktor besitzt als das richtige Protein. Im abschließenden Beispiel wird anhand der Analyse von Spot 8 gezeigt, wie die Darstellungsweise der Software dem Benutzer die Möglichkeit offeriert, Fehler der Peakerkennungssoftware (siehe 3.2.1) zu evaluieren. In Abbildung 42-A ist das Massenspektrum der tryptisch Spaltpeptide des Proteins „*Aconitase Hydratase, Mitchondrial Precursor (EC 4.2.1.3)*" gezeigt.

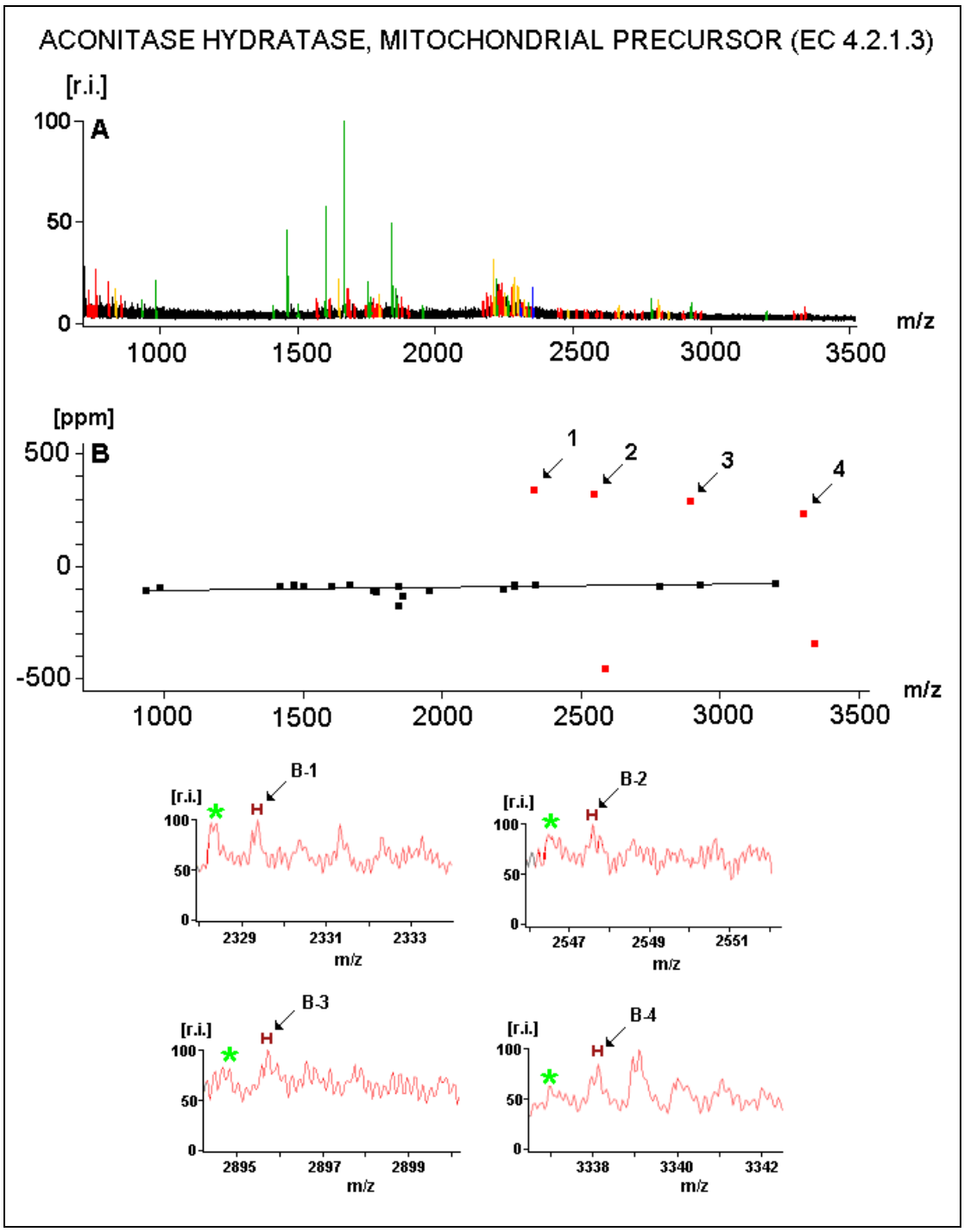

Abbildung 42 Beispiel für falsch selektierte Isotope bei der Analyse des Proteins *„Aconitase Hydratase, Mitchondrial Precursor (EC 4.2.1.3)"*.

In Abbildung 42-B sind 6, von der Software als Ausreißer erkannte Peptidmassen rot eingefärbt, wovon 4 mit Pfeilen markiert sind.

Durch die oben bereits erwähnten interaktiven Eigenschaften sind die jeweiligen Peaks im Spektrum leicht zu lokalisieren. In den 4 Abbildungen unterhalb von Abbildung 42-B sind für die vier mit Pfeilen markierten Ausreißer die Isotopenmuster der entsprechenden Peaks vergrößert dargestellt. In allen vier Fällen wurde das falsche Isotop ausgewählt (durch braun gefärbtes „H" gekennzeichnet). Das richtige, in diesen Fällen durch die Peakerkennungssoftware nicht selektierte Isotop ist mit einem grünen Stern markiert. In Abbildung 43 ist die Streuung der Abweichungen im Bereich von –200 ppm bis 0 ppm um die, durch einfache lineare Regression berechnete, Gerade gezeigt.

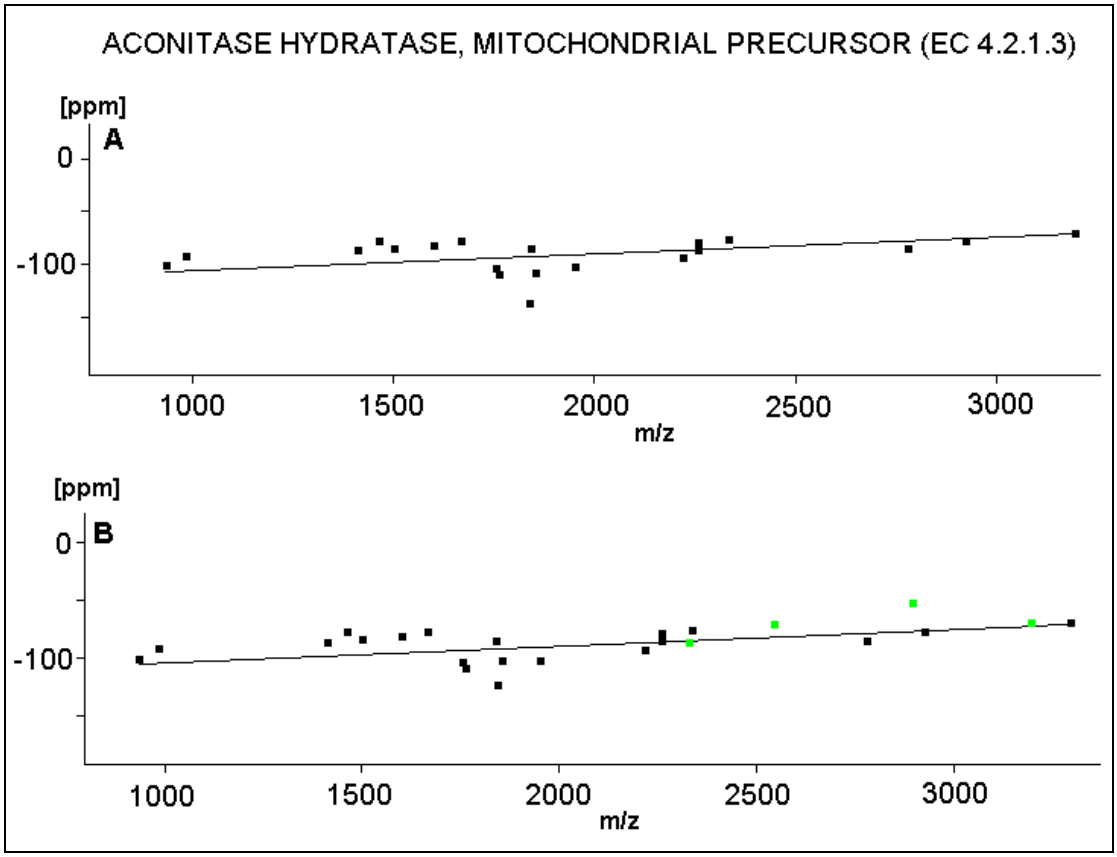

Abbildung 43 Gezeigt ist die Streuung der Abweichungen um die durch lineare Regression berechnete Gerade für das Protein „*Aconitase Hydratase, Mitchondrial Precursor (EC 4.2.1.3)*". (A) Bei insgesamt 4 falsch ausgewählten Isotopen ergeben sich für das Protein 20 Treffer (Modifikationen sind nicht berücksichtigt) und bei richtiger Auswahl aller Isotope werden 4 weitere Treffer (grün markierte Peptidmassen) dem Protein zuerkannt, insgesamt ergeben sich dann 24 Treffer (B).

Abbildung 43-A entspricht Abbildung 42-B nach Eliminierung der dort markierten Ausreißer. Abbildung 43-B zeigt den Fall nach Korrektur des "*Peakpickings*". Nun werden dem Protein weitere 4 Treffer zuerkannt und der Z-Score erhöht sich auf 99,7.