Supplementary Materials for

High-definition *De Novo* Sequencing of Crustacean Hyperglycemic Hormone (CHH)-family Neuropeptides

Chenxi Jia[1], Limei Hui[1], Weifeng Cao[1], Christopher B. Lietz[1], Xiaoyue Jiang[1], Ruibing Chen[1], Adam D. Catherman[2], Paul M. Thomas[2], Ying Ge[3], Neil L. Kelleher[2], and Lingjun Li[1,*]

[1]School of Pharmacy and Department of Chemistry, University of Wisconsin–Madison, Madison, WI 53705

[2]Departments of Chemistry and Molecular Biosciences and the Feinberg School of Medicine, Northwestern University, Evanston, IL 60208

[3]Human Proteomics Program and Department of Cell and Regenerative Biology, University of Wisconsin–Madison, Madison, WI 53706

***Correspondence:** Professor Lingjun Li, School of Pharmacy and Department of Chemistry, University of Wisconsin–Madison, 777 Highland Avenue, Madison, Wisconsin 53705-2222

E-mail: lli@pharmacy.wisc.edu

Phone: +1-608-265-8491

Fax:+1-608-262-5345

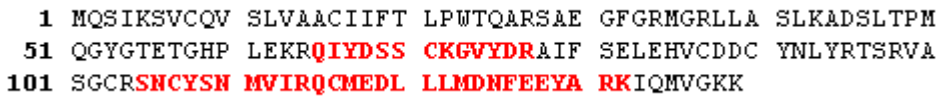

Figure S1. Homologous preprohormone of *P. trituberculatus* CHH hit by Mascot search. Matched peptides are shown in bold red.

Reference: Zhu D.F., Shen J.M., Yang J.F., Su Q. Cloning and characterization of hyperglycemic hormone from the crab *Portunus trituberculatus*. Submitted (JAN-2008) to the EMBL/GenBank/DDBJ databases.

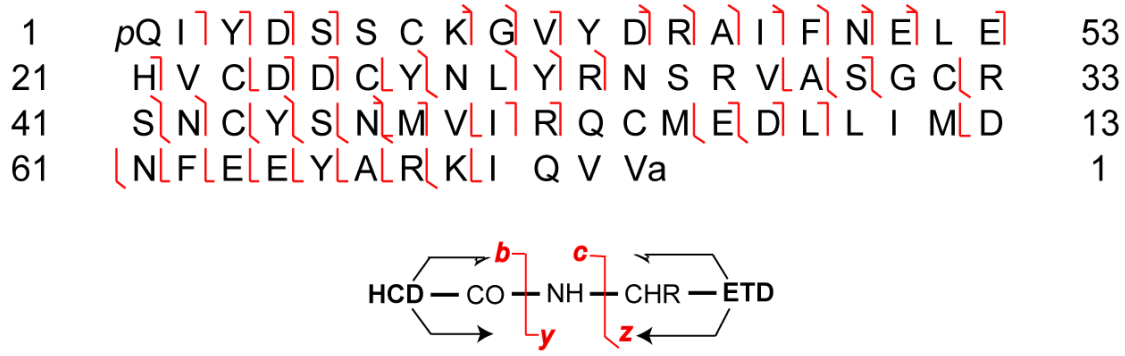

Figure S2. Fragmentation map of *Cas*-SG-CHH by HCD and ETD.

Figure S3. Homologous *Cap*-SG-MIH hit by Mascot search. Matched peptides are shown in bold red.

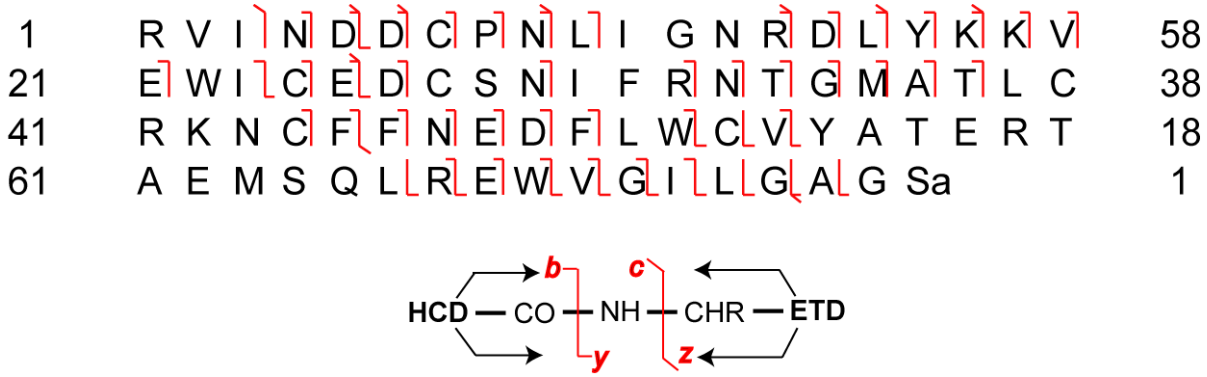

Figure S4. Fragmentation map of $^{Cab\text{-}SG}$MIH by HCD and ETD.

Figure S5. Homologous *Cap*-SG-MOIH hit by Mascot search. Matched peptides are shown in bold red.

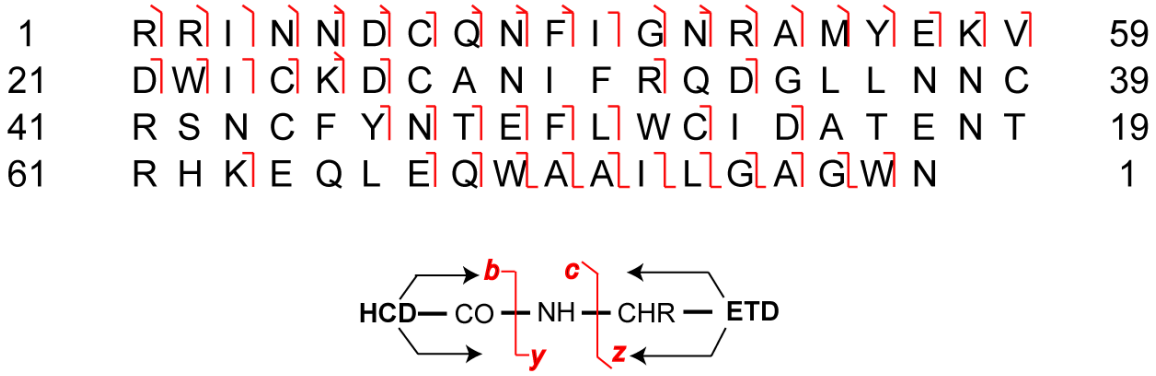

Figure S6. Fragmentation maps of $^{Cab\text{-}SG\text{-}}$MOIH by HCD and ETD.

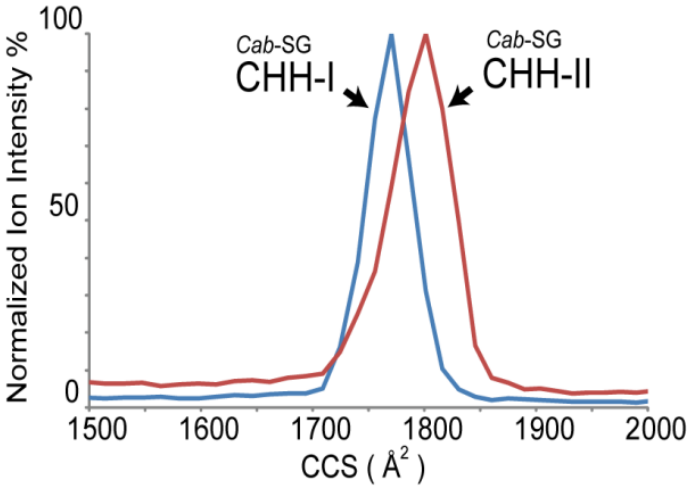

Figure S7. Plot of normalized ion intensity *vs.* collision cross section of $^{Cab\text{-}SG}$CHH-I and $^{Cab\text{-}SG}$CHH-II from ion mobility MS analysis.

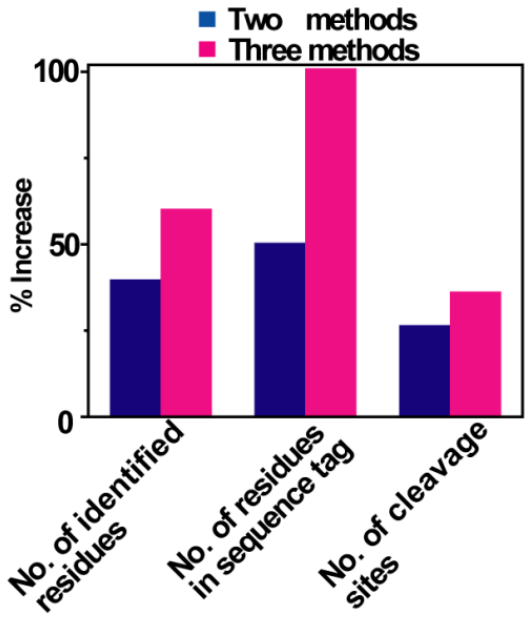

Figure S8. Percent increase of the three numbers as evaluation metrics by combination of two and three methods *vs.* one method. One method: on-line top-down; two methods: on-line top-down and off-line top-down; three methods: on-line top-down, off-line top-down and bottom up.

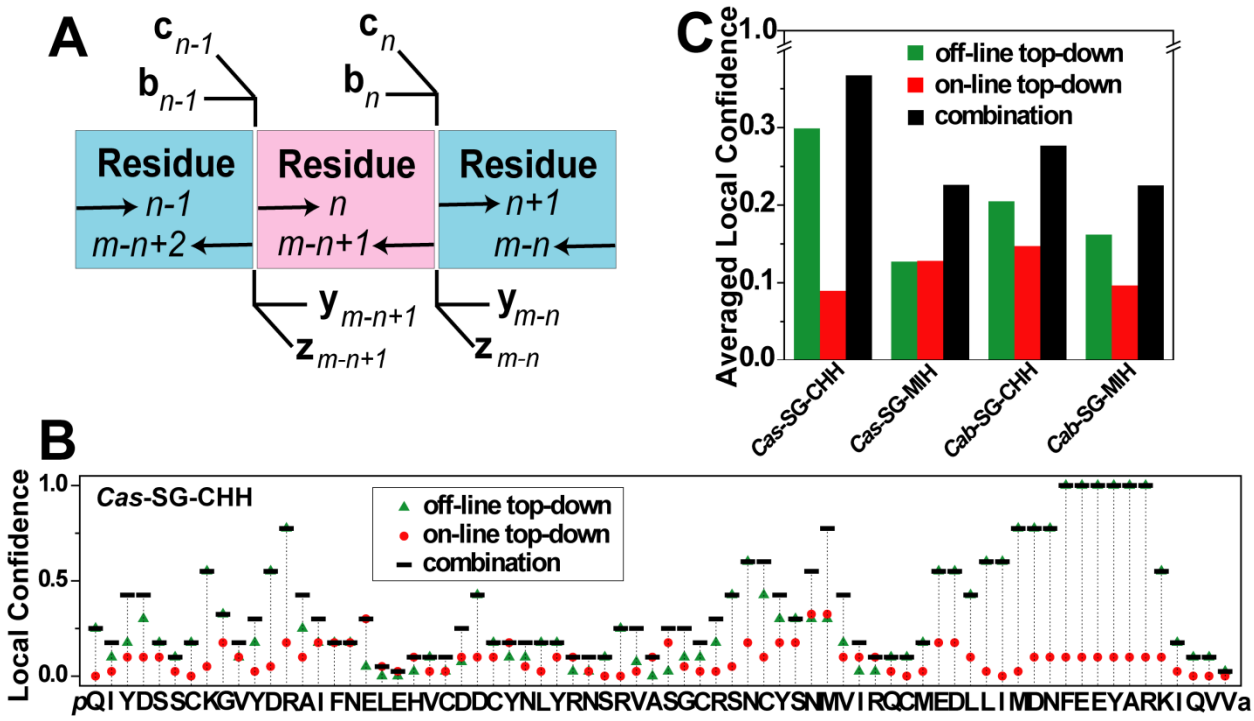

Figure S9. Statistical evaluation of off-line top-down and on-line top-down *de novo* sequencing strategies. (A) Demonstration of ion pairs (same type of two consecutive ions, e.g. b_{n-1}/b_n or y_{m-n+1}/y_{m-n}, etc) and fragment ions for the calculation of local confidence score for the n^{th} residue (counting from N-terminus) of a peptide with a length of *m* residues. (B) Distribution of local confidence scores for *Cas*-SG-CHH. (C) Averaged local confidence scores of *Cas*-SG-CHH, *Cas*-SG-MIH, *Cab*-SG-CHH and *Cab*-SG-MIH. We assume that an ideal identification of the n^{th} residue requires observation of at least eight fragment ions under proton-driven and electron-driven fragmentation, *i.e.* b_{n-1}, b_n, c_{n-1}, c_n, y_{m-n+1}, y_{m-n}, z_{m-n+1}, z_{m-n}. However, in practice, not all these fragments can be detected. Therefore, we define a local confidence score to evaluate the degree of identification for a residue. The local confidence score for a given residue (e.g. the n^{th} residue) is calculated by the following point-based method. 2 points are assigned to an ion pair and 1 point is given to a fragment ion. The local confidence score of the n^{th} residue is calculated by the equation $S_n = (X_{obsd} \cdot 2 + Y_{obsd} \cdot 1)/(X_{thel} \cdot 2 + Y_{thel} \cdot 1)$, where X_{obsd} stands for the total number of observed ion pairs, Y_{obsd} is the total number of observed fragment ions, X_{thel} is the total number of theoretical ion pairs, and Y_{thel} represents the total number of theoretical ions. The averaged local confidence score $\bar{S}$was calculated by the equation $\bar{S} = \sum_1^m S_n/m$, where *m* stands for the total residue number of a peptide.

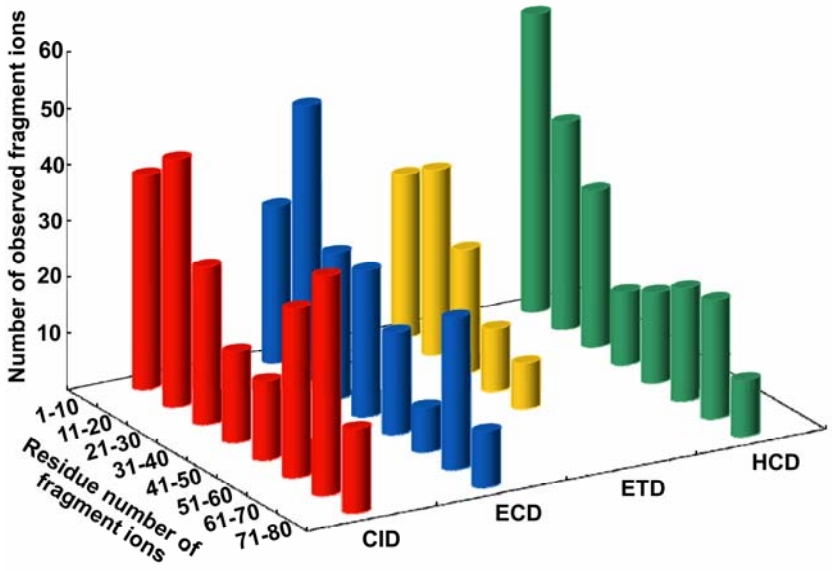

Figure S10. Correlation between length of peptide fragment ions and fragmentation techniques.

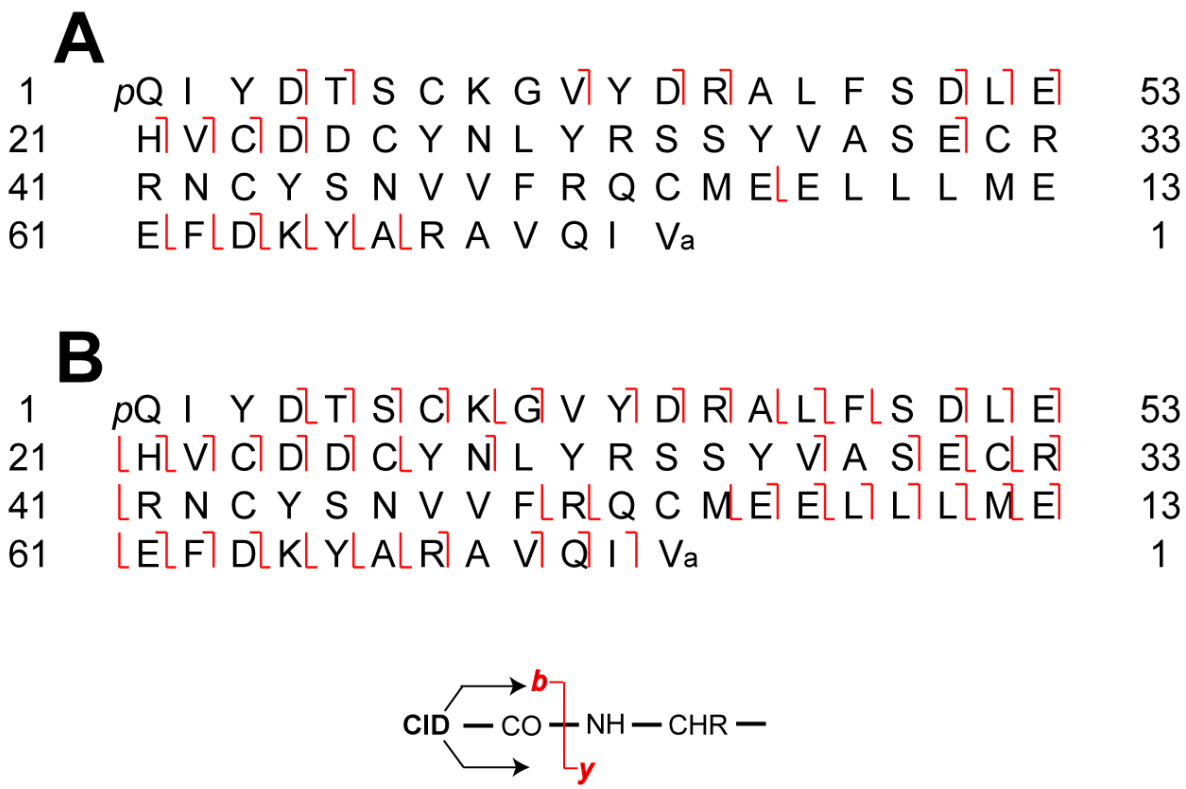

Figure S11. Fragmentation maps of *Cab*-SG-CHH. (A) Our previous study using DTT reduction followed by IAA alkylation. (B) This study using DTT reduction and acidic solvent storage. Compared with our previous study, the current strategy resulted in ~3 fold increase of observed fragment ions and identified sequence coverage.

Table S1. Tryptic peptides of *Cas*-SG-CHH identified by bottom-up *de novo* sequencing

Obsd. *m/z*	Charge state	Calcd MW	Sequence[a]	Tryptic Peptide	Mascot Ion Score[b]
787.30	2	1572.68	*p*QIYDSSCKGVYDR	CHH[1-13]	88
1165.90	2	2330.00	AIFNELEHVCDDCYNLYR	CHH[14-31]	118
622.17	2	1242.55	SNCYSNMVIR	CHH[41-50]	61
1080.38	2	2158.90	*p*QCMEDLLIMDNFEEYAR	CHH[51-67]	62
1088.84	2	2175.92	QCMEDLLIMDNFEEYAR	CHH[51-67]	99

[a]Fixed carbamidomethyl is on cysteine; *p*Q, pyro-Glu.

[b]Matching to homologous sequence.

Table S2. Observed *m/z* and charge state of CHH-family neuropeptides

Species	Peptide name	Obsd. *m/z*	Charge state
C. sapidus	*Cas*-SG-CHH	1211.429	7
	Cas-SG-MIH	1296.064	7
	Cas-PO-CHH	1200.385	7
C. borealis	*Cab*-SG-CHH-I	1220.992	7
	Cab-SG-CHH-II	1223.424	7
	Cab-SG-MIH	1277.051	7
	Cab-SG-MOIH	1322.772	7

Statistical Analysis Evaluates Performance Characteristics of the Three *De Novo* Sequencing Strategies

Complementary Characteristics for Complete Sequence Interpretation. Implementation of three *de novo* sequencing methods provides more confident and effective sequence elucidation of large peptides. Here, we use *Cas*-SG-CHH, *Cas*-SG-MIH, *Cab*-SG-CHH and *Cab*-SG-MIH as model peptides to evaluate the three figures of merit: number of identified residues, number of residues in sequence tags, and number of cleavage sites.

The ultimate goal of *de novo* sequencing is to determine the complete amino acid sequence of a given peptide or protein. The Venn diagram in Figure 9A shows the contributions of the three methods to the number of identified residues. While the four model peptides contain a combined total of 274 residues, only 85 residues are commonly identified by all of the three methods. Each of the three methods enables determination of numerous uniquely identified residues (off-line top-down method yields 20 residues, on-line top-down method shows 11 and bottom-up technique generates 35 residues), and a large number of residues are in the overlap in two methods. The bottom-up method yielded the highest identification percentage, showing11% and 22% higher than the off-line and on-line top-down methods, respectively.

Interpretation of consecutive fragment ions can generate sequence tags, which is critical for peptide sequencing in the absence of a sequenced genome. The Venn diagram in Figure 9B shows the contributions of the three methods to the number of residues that sequence tags contain. The bottom-up method shows the most significant contribution of 65 residues,

indicating that bottom-up MS/MS of tryptic peptides exhibit preference to generate consecutive fragment ions for sequence tag interpretation in comparison with top-down MS/MS of intact large peptides.

The numbers of cleavage sites between two adjacent residues were evaluated. Cleavage site includes both amide bond cleavage and α-carbon bond cleavage primarily produced by proton-driven (CID and HCD) and electron-driven (ECD and ETD) dissociations, respectively. The Venn diagram in Figure 9C illustrates that of the 302 cleavage sites observed, 113 are shared by the three methods.

Figure S8 summarizes the impact of using multiple *de novo* sequencing methods. Combined use of the three methods leads to 60%, 100% and 36% of boosts in the values of the three evaluated aspects, respectively. These results indicate that the three *de novo* sequencing methods have complementary effects on improving sequence coverage.

Local Identification Confidence for Individual Amino Acid Residue. The combination of off-line and on-line top-down methods can also improve the local identification confidence for each amino acid residue. The *de novo* sequencing algorithm PEAKS has a scoring function for evaluation of the local confidence (Ref. 48 in paper). Similarly, we established a simpler evaluation scheme as shown in Figure S9 and legend. Figure S9B plots the confidence scores of each residue of $^{Cas\text{-}SG\text{-}}$CHH. After combination of the two methods, most of scores for individual residues were increased. Figure S9C summarizes the average scores of off-line top-down, on-line top-down and combination for the four model peptides, $^{Cas\text{-}SG\text{-}}$CHH, $^{Cas\text{-}SG\text{-}}$MIH, $^{Cab\text{-}SG\text{-}}$CHH and $^{Cab\text{-}SG\text{-}}$MIH. All the confidence scores are increased with combination of the two methods.

The use of multiple fragmentation techniques, including ECD, ETD, CID, and HCD, produces fragment ions with broader mass ranges in top-down fragmentation of large peptides. Each fragmentation technique has a unique and complementary mass range preference. Figure S10 plots the correlation between length of peptide fragment ions and fragmentation techniques. CID gives most fragment ions located in low and high mass ranges, containing 20 and 70 residues. HCD and ETD give predominantly fragment ions covering the low mass range. ECD produces the fragment ions spanning a broad mass range.

EXPERIMENTAL PROCEDURES (Supplemental Materials)

Animals, Tissue Dissection and Extraction

C. sapidus and *C. borealis* were shipped from the Fresh Lobster Company (Gloucester, MA), and then maintained without food in artificial seawater at 10–12 °C. For tissue collection, animals were anesthetized by packing them in ice for 30–60 minutes. After anesthesia, the eyestalks and dorsolateral portions of the pericardial chamber were removed. The sinus glands (SGs) and pericardial organs (POs) were subsequently dissected from these structures, respectively, in chilled (approximately 4°C) physiological saline (composition: 440 mM NaCl; 11 mM KCl; 13 mM $CaCl_2$; 26 mM $MgCl_2$; 10 mM HEPES acid, pH 7.4, adjusted with NaOH). After dissection, tissue was either pinned in a Sylgard 184 (KR Anderson, Santa Clara, CA)-lined Petri dish and subsequently processed for anatomical studies or immediately placed in acidified methanol (90% methanol:9% glacial acetic acid:1% water) and stored at −80°C.

The tissue was then homogenized, after which the extraction liquid was transferred to a 1.5 ml microcentrifuge tube (Fisher Scientific) and centrifuged at 16,100 rcf for 10 min in an Eppendorf 5415 D microcentrifuge (Brinkmann Instruments). The supernatant fraction was retrieved, and the pellet was re-extracted with acidified methanol and re-centrifuged. Supernatant fractions were combined and concentrated to dryness using a Savant SC 110 SpeedVac concentrator (Thermo Electron). Finally, 100 µl of Millipore water was added to the extract. This resuspended extract was then vortexed and centrifuged. After centrifugation, the clear solution at the upper layer of the extract was used for further analysis.

Multiple Protease Digestion

Three μL of peptide fraction was reduced and alkylated by incubation in 2.5 mM DTT for 1 h at 37 °C followed by incubation in 7 mM iodoacetamide (IAA) in the dark at room temperature for 1 h. A 1 μL of peptide was digested at 37 °C overnight after addition of 50mM ammonium bicarbonate with 0.5 μg of trypsin (Promega, Madison, WI). A 1 μL of peptide was digested overnight at 37 °C after addition of 50 mM Tris-HCl, pH 8.0, and 0.5 μg of Lys-C (Princeton Separations, Adelphia, NJ). One μL of peptide was digested in 2 M urea, 50 mM Tris, 25 mM ammonium bicarbonate at room temperature overnight with 0.5 μg of Glu-C (Princeton Separations, Adelphia, NJ). Finally, each digest was quenched by the addition of formic acid to a final concentration of 0.5%, desalted on a C18 ZipTip (Millipore, Bedford, MA), and the eluent was lyophilized.

Off-line Top-down MS on ESI-UHR-QTOF maXis

Peptide sample was directly infused on a hybrid UHR-QTOF maXis (Bruker Daltonics, Bremen, Germany) by a 25μL Hamilton syringe using infusion pump (KD Scientific, MA, USA). Further settings: flow rate 1.0 μL/min, nebulizer 1.6 psi, dry gas rate 4 L/min, dry gas temperature 200 °C, funnel rf 400 Vpp, multipole rf 400 Vpp, ion energy 5 V, low mass cutoff 300, collision energy 10 V, collision rf 1200 Vpp, ion cooler rf 400 Vpp, transfer time 120 μs, prepulse storage time 10 μs. For top-down MS/MS, the parent ion was selected with a 10 Da isolation window and fragmented by 40% of normalized collision energy. Spectra were acquired

in profile mode with a scan speed of 1.0 Hz. DataAnalysis software (Bruker Daltonics) was used for data processing.

Ion Mobility MS

All ESI-ion mobility (IM)-MS experiments were performed using a Synapt G2 HDMS mass spectrometer equipped with a nano-ESI ion source and MassLynx data processor (Waters, Milford, MA, USA). Instrument acquisition parameters used were as follows: an inlet capillary voltage of 3.0 kV, a sampling cone setting of 35 V, and a source temperature of 120 °C. The argon pressure in the traveling wave ion guide trap and the traveling wave ion guide transfer were 2.44×10^{-2} and 2.61×10^{-2} mbar, respectively. The wave height, the wave velocity, and the nitrogen pressure in the traveling wave IM drift cell were 32.0 V, 850 m/s, and 2.96 mbar, respectively. Samples were directly infused into the mass spectrometer at a rate of 0.5–0.8 μL/min. All IM-MS acquisitions were acquired for 5 min. Data processing was conducted using Waters MassLynx 4.1 and DriftScope 2.1. The experimental procedure for calibration of CCS was described in Supplementary Materials.

Calibration of Collision Cross Section on Ion Mobility MS

The calibration of the SYNAPT G2 travelling wave (TW) IM mass spectrometer to measure gas-phase CCS mainly followed the protocols outlined by Smith et al.[i] The CCS (Ω) of an ion can be calculated as a function of charge (z), ion mass (m_I), drift gas mass (m_N), temperature (T), pressure (P), drift gas number density (N), and drift time (t_D):

$$\Omega = \left[\frac{18\pi}{K_b T}\left(\frac{1}{m_i} + \frac{1}{m_N}\right)\right]^{1/2} \frac{ze}{16}\frac{760}{P}\frac{T}{273.2}\frac{1}{N}At_D{}^B \qquad (1)$$

The constants in Equation 1 include the Boltzmann constant (K_b), elementary charge (e), a correction factor for the pulsed electric field parameters (A), and an exponent to compensate for the non-linear Ω-t_D relationship in a TW-IMS instrument (B). To simplify the calculation, equation 1 can be divided by ze and reduced mass to yield a mass and charge independent CCS (Ω'). Then, all constants are combined with A to yield a new constant (A') and produce a new equation:

$$\Omega' = A' t_D{}^B \quad (2)$$

A set of standards were analyzed with known CCS values[ii] to obtain values for A' and B. The Ω' of an ion can be determined by plugging the drift time into Equation 2, and Ω can be calculated as the product of Ω', the absolute charge of ion, and the square-root of the inverse reduced mass.

Table S2. Ions used for CCS calibration.

Standard	z	mass (Da)	Ω (nm^2)
Ubiquitin	7	8561	19.1
Ubiquitin	8	8561	19.9
Ubiquitin	9	8561	20.9
Ubiquitin	10	8561	22.0
Ubiquitin	11	8561	23.4
Ubiquitin	12	8561	24.8
Ubiquitin	13	8561	26.0
Myoglobin	15	16951	40.6
Myoglobin	16	16951	41.8
Myoglobin	17	16951	43.1
Myoglobin	18	16951	44.4
Myoglobin	19	16951	45.7
Myoglobin	20	16951	47.0
Myoglobin	21	16951	48.2
Myoglobin	22	16951	49.2
Myoglobin	23	16951	50.1
Myoglobin	24	16951	50.9
Cytochrome C	13	12327	30.8
Cytochrome C	14	12327	32.0
Cytochrome C	15	12327	33.3
Cytochrome C	16	12327	34.5
Cytochrome C	17	12327	36.0
GRGDS	1	490.21	2.06
GRGDS	2	490.21	2.15
Angiotensin Fragment	2	898.47	3.32
RASG-1	2	1000.49	3.32
Angiotensin II	2	1045.53	3.74
Bradykinin	2	1059.56	3.67
Angiotensin I	3	1295.68	3.12
Renin Substrate	3	1757.93	3.88
Enolase T35	3	1871.96	4.16

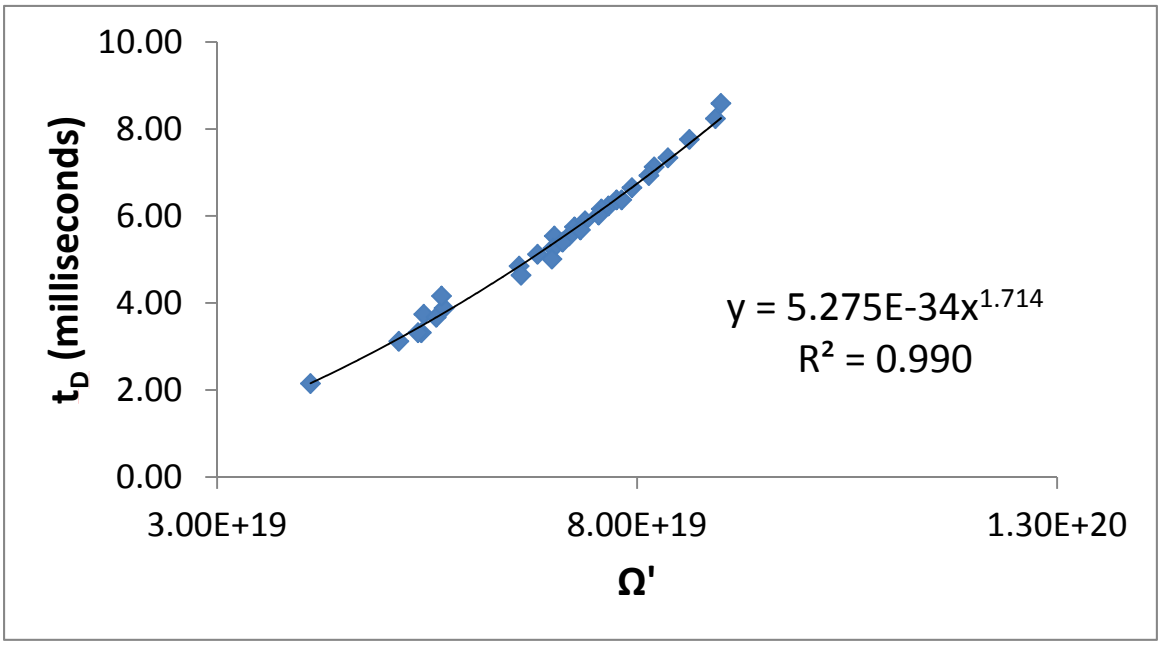

Figure S12. Calibration curve of Ω' *vs.* t_D.

[i] Smith, D. P.; Knapman, T. W.; Campuzano, I.; Malham, R. W.; Berryman, J. T.; Radford, S. E.; Ashcroft, A. E. Deciphering drift time measurements from travelling wave ion mobility spectrometry-mass spectrometry studies. *Eur. J. Mass Spectrom.***2009**, *15*, 113-130.

[ii] Bush, M.F.; Hall, Z.; Giles, K.; Hoyes, J.; Robinson, C. V.; Ruotolo, B. T. Collision cross sections of proteins and their complexes: a calibration framework and database for gas-phase structural biology. *Anal. Chem.***2010**, *82*, 9557-9565.