Peptide Fragmentation

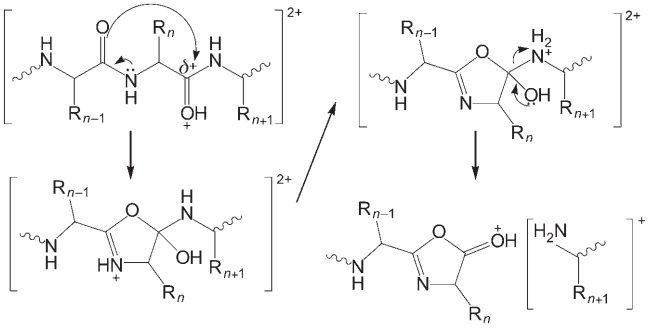
DOI: 10.1002/anie.200603881

Backbone Carbonyl Group Basicities Are Related to Gas-Phase Fragmentation of Peptides and Protein Folding**

*Mikhail M. Savitski, Frank Kjeldsen, Michael L. Nielsen, Sergiy O. Garbuzynskiy, Oxana V. Galzitskaya, Alexey K. Surin, and Roman A. Zubarev**

Herein, we demonstrate that the same fundamental parameter lies behind such disparate phenomena as folding of proteins in solutions and fragmentation of peptide cations in the gas phase. This parameter is the gas-phase basicity (GB) of backbone carbonyl groups. GB relates to the free energy of accepting a proton, and it is responsible, for example, for protonation of polypeptides in electrospray ionization (ESI).[1] GB values of free amino acids and short peptides have been measured experimentally,[2,3] but these studies could not separate the effects of other groups from the basicity of the backbone carbonyl groups. Zhang provided the first GB estimates of backbone amides[4] from a kinetic model of peptide fragmentation. The model was trained on mass spectrometry (MS) data obtained in proteomics research,[5–9] a valuable source for fragmentation studies.[10–12] However, the kinetic model did not separate basicities of backbone carbonyl groups from those of the NH groups.[4,13] Herein, we derive carbonyl group GB values by using a different model of fragmentation and larger statistics.

The generally accepted peptide-fragmentation mechanism (Scheme 1) is based on proton mobilization onto the backbone carbonyl group oxygen of the *n*th residue with subsequent attack by the $(n-1)$th carbonyl group oxygen center on the partially electropositive carbon atom of the protonated carbonyl group.[14,15] Subsequent proton transfer to the nitrogen atom results in C–N bond cleavage (Scheme 1).[16] Similar to the kinetic model,[4,13] the rate of intramolecular proton transfer is considered faster than the bond rupture. Protons are assumed to be statistically distributed over backbone carbonyl groups according to their basicities. The rate constant of the proton transfer to the nitrogen atom is the same for all amino acids. The model predicts the cleavage probability to be determined by the



Scheme 1. Generally accepted peptide-fragmentation mechanism.

frequency of protonation of the *n*th carbonyl group, that is, by the carbonyl group basicity of the *n*th residue. If the carbonyl group is engaged in long-lived hydrogen bonding, its ability to accommodate additional protons will be reduced. However, in tryptic peptides that are 10–12-residues long and activated during the collisionally activated dissociation (CAD) process to 200–400 °C,[17] neutral hydrogen bonding is relatively short-lived and does not cause major disruptions of intramolecular proton transfer.

The predictions of the carbonyl group basicity model was tested on dications of tryptic peptides, the most abundant ionic species in ESI-based proteomics.[18] In these species, one charge is sequestered at the C-terminal Lys or Arg residue, whereas the second mobile charge is located close to the N terminus.[19] Amino acids Arg, Lys, and Trp, which are rarely found in internal parts of tryptic peptides, were not considered. Cysteine was also excluded as its side chain is usually alkylated prior to MS analysis. The remaining 16 amino acids were separated into a core group (Ala, Gly, Phe, Ile, Leu, Met, Pro, Ser, Thr, and Val) and a special group of residues whose side chains form hydrogen bonds with their own amides,[3] that is, Glu, Asp, Gln, Asn, and His. Side chains of Glu and Asp can donate a proton to the carbonyl group, which enhances cleavages after these residues, especially after Asp.[20] Gln and Asn are known to promote NH_3 losses,[21] which can involve cyclization interfering with C–N bond cleavage. The high basicity of His can obstruct intramolecular proton flow as the His side chain can capture the mobile proton and then donate it to its backbone carbonyl group.[14,22]

As formation of b_1 and b_2 ions are special cases,[16,23] CAD statistics only included cleavages leading to $y_{k-3} \rightarrow y_{k-7}$ fragments (k is the peptide length) and the complementary $b_3 \rightarrow b_7$ ions. Cleavage propensity for each amino acid was calculated, similar to that shown in reference [12], as the relative frequency of cases when cleavage after the amino acid

[*] M. M. Savitski, Dr. F. Kjeldsen, M. L. Nielsen, Prof. R. A. Zubarev
Laboratory for Biological and Medical Mass Spectrometry
BMC, Uppsala University
Box 583, 75123 Uppsala (Sweden)
Fax: (+46) 18-471-22-44
E-mail: roman.zubarev@bmms.uu.se
S. O. Garbuzynskiy, O. V. Galzitskaya, A. K. Surin
Institute of Protein Research
Russian Academy of Sciences
142290, Pushchino, Moscow Region (Russia)

[**] This work was supported by the Knut and Alice Wallenberg Foundation and Wallenberg Consortium North (grant WCN2003-UU/SLU-009 to R.Z. and instrumental grant to R.Z. and Carol Nilsson) as well as the Swedish research council (grants 621-2004-4897, 621-2002-5025, and 621-2003-4877 to R.Z.). Thomas Köcher and Christopher Adams are acknowledged for insightful discussion.

produces the most abundant *y* ion in the mass spectrum. The calculations were based on a library of 15000 high-resolution CAD MS/MS spectra of tryptic peptides.[12] The reason why cleavage after an amino acid was taken is explained in Figure 1, in which the variance of propensities for 11 "core" amino acids is plotted against their position relative to the cleavage site. A larger variance means a more important position, thus AA- (*n*th position in Scheme 1) is more important than -AA (where - indicates the cleavage site and AA represents an amino acid). This is because *b* ions are less stable than *y* ions,[24] and thus for *b* ions the nature of the terminal side chain is more important.

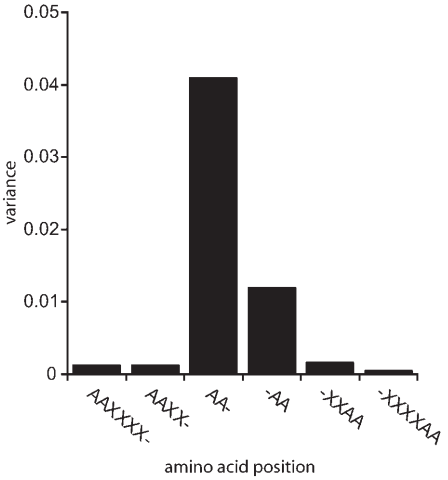


Figure 1. Variance of CAD cleavage propensities as a function of the amino acid position (AA) relative to the cleavage site (-). Position AA- corresponds to position *n* in Scheme 1.

The proof that the average relative cleavage propensities (Table 1) reflect GB of the backbone carbonyl groups came from the analysis of a database[25] containing 3769 protein structures with a total of 690067 residues forming 455300 backbone–backbone H-bonds. The assumption was that carbonyl group basicity should direct formation of backbone–backbone hydrogen bonding in α helices and β sheets. Moreover, the same intrinsic property should determine the participation rate of amino acids in these well-organized secondary structures. For each amino acid, the H-bond-accepting propensity (P_H) was calculated as n_H/n_{tot}, where n_H is the number of hydrogen bonds accepted by the amino acid from other backbone amides and n_{tot} is the amino acid occurrence in the database. The structure-forming propensity (P_S) was defined as n_{AA}/n_{tot}, where n_{AA} is the amino acid occurrence in α helices and β sheets. Data for 16 amino acids are summarized in Table 1.

Although P_S and P_H are independently obtained parameters, a strong correlation ($r = 0.94$) was found between them, indicating that the same intrinsic property determines both formation of backbone–backbone H-bonding and participation in well-organized structures. As the main common factor is the H-bond acceptance of amino acids, this intrinsic property must be the backbone carbonyl group basicity.

Table 1: Relative propensity P_{CAD} to cleavage in CAD of peptide bond after an amino acid and crystal-structure data.

AA	P_{CAD}	n_{AA}[a]	n_{tot}[b]	n_H[c]
Ala	0.59	37677	56123	38415
Asp	0.63	17770	40481	25232
Glu	0.69	30187	46658	29332
Phe	0.60	18593	27585	19959
Gly	0.24	15771	49881	25020
His	0.68	8683	15770	10383
Ile	0.74	29589	39442	31098
Leu	0.71	45209	63270	47115
Met	0.65	10068	15183	10851
Asn	0.45	12623	30385	17616
Pro	0.13	9277	31964	15236
Gln	0.67	16680	26457	16755
Ser	0.41	19863	41072	24432
Thr	0.52	20605	38084	24258
Val	0.75	35988	49062	37793
Tyr	0.58	16079	23943	17509

[a] n_{AA} = number of residues found in α helices and β sheets; [b] n_{tot} = total number of residues in the database; [c] n_H = number of carbonyl-group-accepted hydrogen bonds with backbone amides found in the whole protein database.

Table 2 shows the relative GB values for 16 amino acids evaluated from the best fit between P_S and P_H. Now the hypothesis of GB directing CAD cleavage can be easily tested.

Table 2: Relative and absolute gas-phase basicities (GB_{rel} and GB, respectively) of backbone carbonyl groups evaluated from crystal-structure data.

AA	GB_{rel}	GB[a] [kcal mol^{-1}]
Ala	0.69	207.3
Asp	0.62	206.7
Glu	0.68	207.3
Phe	0.76	208.0
Gly	0.12	**202.7**
His	0.71	207.5
Ile	0.98	**210.8**
Leu	0.87	**209.6**
Met	0.78	208.2
Asn	0.41	204.9
Pro	0	201.3
Gln	0.66	207.1
Ser	0.40	204.8
Thr	0.55	206.2
Val	0.95	**208.7**
Tyr	0.76	208.0

[a] Absolute basicity was found from scaling by using the best linear fit ($r = 0.96$) to the reference GB data[3] for Gly, Ala, Val, Leu, and Ile (highlighted in boldface). The order of the data is more reliable than the absolute values.

Comparison of P_S and P_H with cleavage propensities for 11 amino acids is shown in Figure 2. In both cases, excellent correlation is found ($r = 0.98$). Not surprisingly, correlation with evaluated data from Table 2 is even higher, $r = 0.984$ (not shown). Note that neither Pro nor Gly, usually considered to be special cases,[11,15,20] are outliers in Figure 2. The impact of

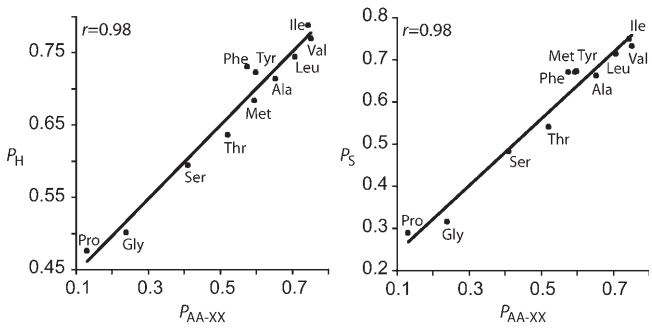


Figure 2. Comparison of P_H and P_S values with CAD cleavage frequencies for 11 core-group amino acids (Table 1). P_{AA-XX} represents the propensity for AA–XX cleavage in CAD.

side chains of the special-group amino acids Asn, Asp, Gln, Glu, and His in the AA- position on frequencies of CAD cleavages can now be estimated from the comparison of their relative basicity data and CAD cleavage frequencies.

In Figure 3, the corresponding plot is presented with the trend (solid line) determined by the 11 core-group amino

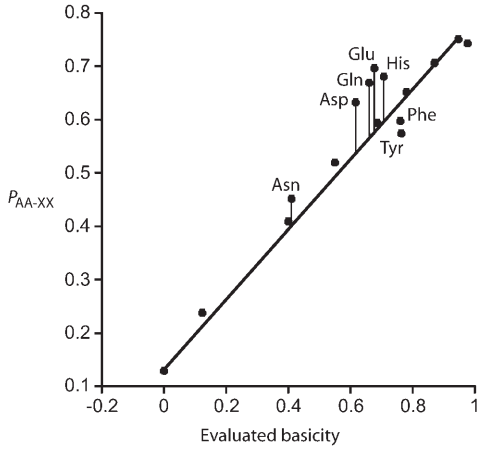


Figure 3. Estimation of the side-chain effect on fragmentation behavior of the amino acids Asn, Asp, Gln, Glu, and His in the *n*th position. Data from Table 2 is correlated with CAD cleavage propensities. The main trend (solid line) is determined for 11 core-group amino acids.

acids. Deviation of the top five residues from the general trend is the effect of their side chains, with the impact for Glu being the highest and that for Asn being the lowest. That Asp scored fairly modestly was no surprise: the Asp side chain promotes CAD cleavages mainly in the absence of labile protons.[20] Note that for all five deviating amino acids, the side-chain effect on cleavage propensity was smaller than the effect of carbonyl group basicity and that only His propensity deviates in this group from the basicity order.

Thus, the hypothesis for carbonyl group basicity is largely relevant even for these five deviating residues. The model of carbonyl group basicity can also explain more-subtle phenomena, such as the difference[11] between the cleavage propensities of isomeric residues Leu (GB 0.98) and Ile (GB 0.87). The absolute GB values can be estimated (Table 2, right column) by using a linear fit ($r = 0.96$) between the relative GB values for carbonyl groups and the reference absolute GB values for free amino acids (boldface in Table 2) according to Equation (1).

$$GB_{abs} = 8.738\,GB_{rel} + 201.34\ \text{kcal mol}^{-1} \quad (1)$$

As this approach is focused on carbonyl group basicity and ignores other groups, the actual basicity of an amino acid may deviate from Equation (1) by a few kcal mol^{-1}. Otherwise, values given by Equation (1) look reasonable. The range of these values, 201.3–209.9 kcal mol^{-1}, overlaps with GBs of free amino acids, 202.7–223.7 kcal mol^{-1}.[3] This overlap explains why a mobile proton is rapidly transferred in CAD to backbone carbonyl groups. As already mentioned, carbonyl groups that engage in persistent hydrogen bonding have reduced basicities, which can explain the lower rates of cleavages after such carbonyl groups.[26]

Experimental Section

The database of protein structures was created based on the structural classification of proteins (SCOP) (25) database version 1.65 release. 3769 Proteins with less than 25% sequence identity belonging to SCOP classes *a*, *b*, *c*, and *d* were selected for the analysis: 794 all-α proteins from class *a*, 928 all-β proteins from class *b*, 1089 α/β proteins from class *c*, and 958 α + β proteins from class *d*. The number of backbone–backbone hydrogen bonds was calculated separately for helices (α helices and 3_{10} helices), β strands, and other parts of structures including unstructured regions, loops, and β turns, among others. A standard program, database of secondary structure assignments (DSSP),[27] was used to identify backbone–backbone hydrogen bonds and calculate their energies. H-bonds were defined by using a 0.5 kcal mol^{-1} energy cutoff. The analyzed structures contained 251130 residues in α-helical regions that formed approximately 160000 H-bonds, 152182 residues in β-sheet regions with 109500 H-bonds, and 286755 residues in other regions with 185800 H-bonds in these regions.

Product-moment analysis[28] was employed for determining correlation between *n* pairs of *x* and *y*. The interpretation of *r* values depends upon *n* and the distribution of *x* and *y*. For normal distributions, thresholds for statistical validity are tabulated.[29] The default threshold probability was chosen to be 1%. The threshold value for *r* then was 0.708 for 11 amino acids and 0.590 for 16 amino acids.

Received: September 21, 2006
Revised: November 28, 2006
Published online: January 9, 2007

Keywords: hydrogen bonds · mass spectrometry · peptide fragmentation · protein structures · proteomics

[1] J. B. Fenn, M. Mann, C. K. Meng, S. F. Wong, C. M. Whitehouse, *Science* **1989**, *246*, 64.
[2] J. Y. Wu, C. B. Lebrilla, *J. Am. Chem. Soc.* **1993**, *115*, 3270.
[3] A. G. Harrison, *Mass Spectrom. Rev.* **1997**, *16*, 201.
[4] Z. Q. Zhang, *Anal. Chem.* **2004**, *76*, 3908.
[5] D. F. Hunt, J. R. Yates, J. Shabanowitz, S. Winston, C. R. Hauer, *Proc. Natl. Acad. Sci.* **1986**, *83*, 6233.
[6] J. K. Eng, A. L. McCormack, J. R. Yates, *J. Am. Soc. Mass Spectrom.* **1994**, *5*, 976.

[7] A. Pandey, M. Mann, *Nature* **2000**, *405*, 837.
[8] R. Aebersold, M. Mann, *Nature* **2003**, *422*, 198.
[9] A. I. Nesvizhskii, R. Aebersold, *Drug Discov. Today* **2004**, *9*, 173.
[10] Y. Y. Huang, J. M. Triscari, G. C. Tseng, L. Pasa-Tolic, M. S. Lipton, R. D. Smith, V. H. Wysocki, *Anal. Chem.* **2005**, *77*, 5800.
[11] Y. Y. Huang, J. M. Triscari, L. Pasa-Tolic, G. A. Anderson, M. S. Lipton, R. D. Smith, V. H. Wysocki, *J. Am. Chem. Soc.* **2004**, *126*, 3034.
[12] M. M. Savitski, F. Kjeldsen, M. L. Nielsen, R. A. Zubarev, *Angew. Chem.* **2006**, *32*, 5427; *Angew. Chem. Int. Ed.* **2006**, *45*, 5301.
[13] Z. Q. Zhang, *Anal. Chem.* 2005, 77, 6364.
[14] V. H. Wysocki, G. Tsaprailis, L. L. Smith, L. A. Breci, *J. Mass Spectrom.* **2000**, *35*, 1399.
[15] A. Schlosser, W. D. Lehmann, *J. Mass Spectrom.* **2000**, *35*, 1382.
[16] A. L. McCormack, A. Somogyi, A. R. Dongre, V. H. Wysocki, *Anal. Chem.* **1993**, *65*, 2859.
[17] P. D. Schnier, J. C. Jurchen, E. R. Williams, *Journal of Physical Chemistry B* **1999**, *103*, 737.
[18] M. L. Nielsen, M. M. Savitski, R. A. Zubarev, *Mol. Cell. Proteomics* **2005**, *4*, 835.
[19] A. R. Dongre, A. Somogyi, V. H. Wysocki, *J. Mass Spectrom.* **1996**, *31*, 339.
[20] W. Yu, J. E. Vath, M. C. Huberty, S. A. Martin, *Anal. Chem.* **1993**, *65*, 3015.
[21] D. L. Tabb, L. L. Smith, L. A. Breci, V. H. Wysocki, D. Lin, J. R. Yates, *Anal. Chem.* **2003**, *75*, 1155.
[22] D. L. Tabb, Y. Y. Huang, V. H. Wysocki, J. R. Yates, *Anal. Chem.* **2004**, *76*, 1243.
[23] T. Yalcin, C. Khouw, I. G. Csizmadia, M. R. Peterson, A. G. Harrison, *J. Am. Soc. Mass Spectrom.* **1995**, *6*, 1165.
[24] D. J. Aaserud, D. P. Little, P. B. Oconnor, F. W. McLafferty, *Rapid Commun. Mass Spectrom.* **1995**, *9*, 871.
[25] A. G. Murzin, S. E. Brenner, T. Hubbard, C. Chothia, *J. Mol. Biol.* **1995**, *247*, 536.
[26] Z. Q. Zhang, J. Bordas-Nagy, *J. Am. Soc. Mass Spectrom.* **2006**, *17*, 786.
[27] W. Kabsch, C. Sander, *Biopolymers* **1983**, *22*, 2577.
[28] K. Pearson, *Philos. T. Roy. Soc. A* **1896**, *187*, 253.
[29] V. Bewick, L. Cheek, J. Ball, *Crit. Care* **2003**, *7*, 451.

