Towards quantitative prediction of proteasomal digestion patterns of proteins

Denis S Goldobin[1,2] **and Alexey Zaikin**[3,4]

[1]Department of Physics, University of Potsdam, Postfach 601553, D–14415 Potsdam, Germany

[2]Department of Theoretical Physics, Perm State University, 15 Bukireva str., 614990 Perm, Russia

[3]Department of Mathematical & Biological Sciences, University of Essex, Wivenhoe park, CO4 3SQ Colchester, UK

[4]Departments of Mathematics & Institute of Women Health, University College London, Gower street, WC1E 6BT London, UK

E-mail: `Denis.Goldobin@gmail.com`

Abstract. We discuss the problem of proteasomal degradation of proteins. Though proteasomes are important for all aspects of the cellular metabolism, some details of the physical mechanism of the process remain unknown. We introduce a stochastic model of the proteasomal degradation of proteins, which accounts for the protein translocation and the topology of the positioning of cleavage centers of a proteasome from first principles. For this model we develop the mathematical description based on a master-equation and techniques for reconstruction of the cleavage specificity inherent to proteins and the proteasomal translocation rates, which are a property of the proteasome specie, from mass spectroscopy data on digestion patterns. With these properties determined, one can quantitatively predict digestion patterns for new experimental set-ups. Additionally we design an experimental set-up for a synthetic polypeptide with a periodic sequence of amino acids, which enables especially reliable determination of translocation rates.

PACS numbers: 05.40.-a, 87.15.R-, 87.15.km, 87.19.xw

Special Issue: Article preparation, IOP journals

A macromolecular complex, the proteasome, is the complex molecular machine for the degradation of intracellular proteins [1]. In particular, proteasomes produce epitopes for an immune system [2]. They exist in cells as the free proteolytically active core, the barrel-shaped 20S proteasome (figure 1), and as associations of this core with regulatory complexes PA700 (19S regulator) or PA28 (11S regulator) at its ends [3]. This paper deals with proteasomal digestion of proteins widely studied in molecular biology and immunology.

A protein enters the proteasome and is transported into the central chamber (this process is referred as the *translocation* one) where it is cleaved into fragments by one of the cleavage terminals arranged along two rings. Fragments of the protein produced are removed through proteasome gates. Some of these fragments, epitopes, are transported onto the cell surface where T-lymphocytes scan them in order to recognize the cells to be killed because of an abnormal functioning. Hence, the digestion pattern for a degraded protein and its statistical properties determine the reaction of the immune system to the presence of this protein in a certain cell. Peculiarities of the translocation rates can qualitatively affect the expression of the specific fragment, *e.g.*, an epitope, because an altered transport changes time of being near the cleavage terminal, *i.e.*, conditions of cleavage. Moreover, impairment of proteasomal degradation, probably due to transport malfunction, might contribute to the pathology of various neurodegenerative conditions [4].

The mechanism of protein translocation remains unknown (however, subjects related to some extent to this problem have been studied in [5, 6, 7, 8]). It is also unknown whether this mechanism is qualitatively different for different proteasome types (constitutive or immuno-), with/without different regulatory complexes. Recently, in [9] a stochastic model, which allows a straightforward reconstruction of the translocation rates and cleavage specificities from mass spectroscopy (MS) data on digestion patterns, has been introduced. These properties reconstructed can be used for a comprehensive quantitative prediction of proteasomal digestion patterns for new proteins and new experimental set-ups. In this paper we elaborate the mathematical theories for the employing of the introduced model for relatively *short synthetic polypeptides* (section 2), *long proteins with a periodic sequence of amino acids* (section 3), and *long natural proteins* which require a peculiar approach (section 4).

1. Physical model of the system and mathematical description

We describe the process of protein transport and degradation by the proteasome (see figure 1) within the framework of the following assumptions.

• **Protein translocation:** The process of the infiltration of a protein into the proteasome chamber is a sequence of thermal noise induced jumps of the protein strand by one amino acid (AA). In figure 1, the zoom-in of the chamber gate schematically shows the diameter of the gate to be comparable with the characteristic size of an AA, what means that the protein chain may be fixed in metastable states by a tight gate

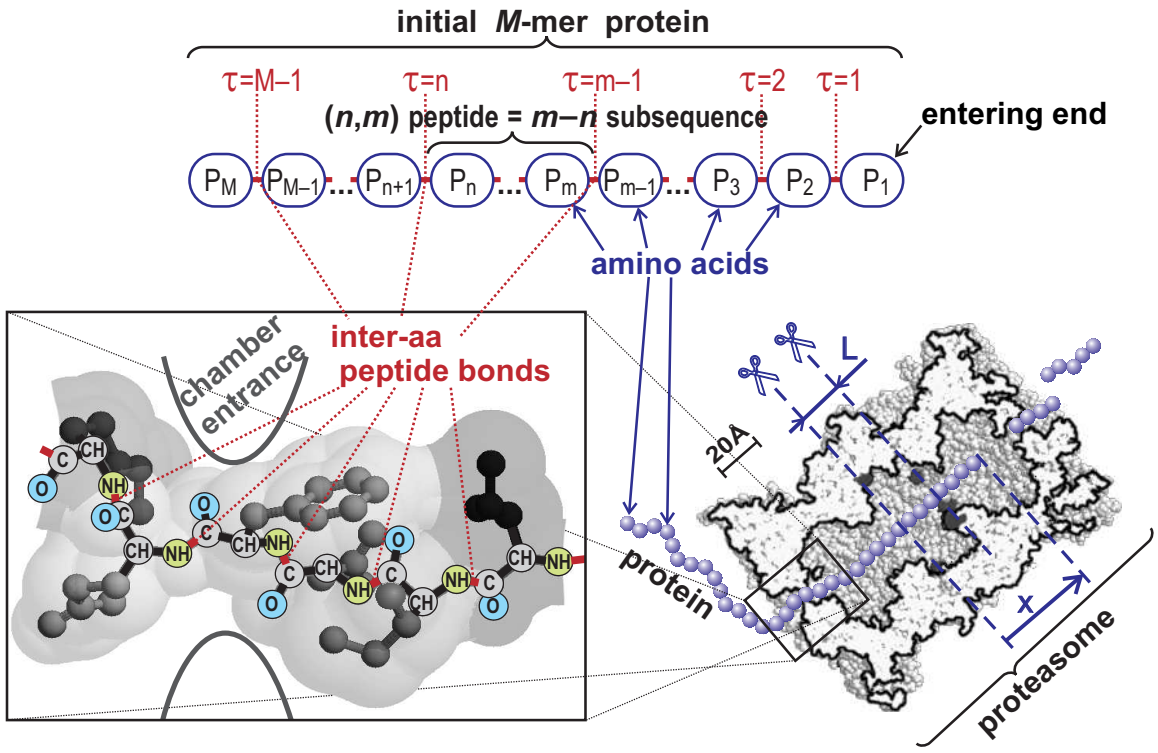

Figure 1. Infiltration of a protein strand into the 20S proteasome: The scissors mark the positions of active sites rings at $x = 0$ and $x = L$; the cleavage occurs via the attaching-detaching of the protein to active sites (dark-grey color). The zoom-in shows the protein fragment `KEFNII` passing through the gate; the electron shields are presented in pale colors.

between successive jumps due to large thermal fluctuations. Indeed, the atomic force measurements reveals $U_b/kT > 3$ [10], where U_b is the characteristic height of the energy barrier separating nearest metastable positions of the chain and $kT/2$ is the energy of thermal fluctuations. The probability of the protein shift by one AA during the infinitesimal time interval dt into the proteasome (to the right in figure 1) is assumed to depend only on the length x of the protein forward end beyond the active sites nearest to the proteasome chamber gate used for protein infiltration (the left ones in figure 1); this probability divided by dt is given by the translocation rate function (TRF) $v(x) \equiv v_x$. In such a way, we neglect the role of the AA sequence specificity for translocation, what is suggested by a non-covalent interaction between the proteasome and the retracted protein. The backward motions of the entering strand are neglected as well (from [10], for the potential energy $U(x)$ of the metastable state x, $(U(x-1)-U(x+1))/2kT \approx 2.5$, thus, meaning the probability of a backward motion to be diminished by factor $\mathrm{e}^{-2.5}$ against the forward one). These assumptions do not impose significant restrictions on the physical mechanism of the translocation process: they are valid for the thermal drift in a tilted spatially-periodic potential (*e.g.*, see [11]) as well as for the ratchet effect (*e.g.*, see [8]), *etc.* The TRFs of different proteasome species (20S, 26S which is the association of 20S core and 19S regulatory complexes, *etc.* [3]) differ.

• **Cleavage:** When the protein strand is close to the active site, the probability of cleavage during the infinitesimal time interval dt depends on the sequence of AAs nearest to the peptide bond cleaved [12]. For the given protein, this conditional probability divided by dt, in other words, conditional cleavage rate (CCR), $\gamma(\tau) \equiv \gamma_\tau$, is a function of the bond number τ (precisely, τ numerates the position of the bond within the initial protein and is counted from the end which has first entered the proteasome; see figure 1). In the following we use the number τ of the bong nearest to the first ring of active sites as a *time-like variable*.

• **Removal of digestion products:** The cleaved parts of the protein degraded, peptides, leave the chamber through the second proteasome gate. Due to their mobility being higher in comparison to that of the protein, processed peptides leave the chamber quick enough to neglect both their possible further splitting and their influence on the protein translocation.

Let us now introduce the distribution $w(x|\tau)$ which is the probability of the protein forward end beyond the first ring of the active sites to be of the length x, when the τth bond is near that ring, in terms we use henceforth, at the discrete "time moment" τ. We measure x in AA. Note, x and τ are integer. In the following we describe the "temporal" evolution of distribution $w(x|\tau)$. On this way, we treat the shift of the protein strand into the proteasome for one AA, *i.e.*, the transition $\tau \to \tau + 1$. Let us decompose $w(x|\tau + 1)$ as

$$w(x|\tau + 1) = \textstyle\sum_j w_j(x|\tau + 1)\,,$$

where $w_j(x|\tau + 1)$ are the contributions due to different scenarios of this transition. Along with $w(x|\tau)$, we account $Q(n, m|\tau)$, the amount of the peptide (n, m), which is the m–n subsequence of the degraded protein (see figure 1), generated during transition $\tau \to \tau + 1$.

In the process of protein digestion there are three possible elementary events:
(a) the strand shift: $x \to x + 1$, $\tau \to \tau + 1$; the event rate is $v(x)$;
(b) the cleavage on the first ring of cleavage centers ($x = 0$): $x \to 0$, $\tau \to \tau$; the event rate is $\gamma(\tau)$;
(c) the cleavage on the second ring of cleavage centers ($x = L$, L is the distance between the rings of cleavage centers): $x \to L$, $\tau \to \tau$; the event rate is $\gamma(\tau - L)$.

In terms of these elementary events the possible scenarios of transition $\tau \to \tau+1$ are
(1) Elementary event (a). Its probability is

$$P_1(x|\tau) = \begin{cases} v_x/(v_x + \gamma_\tau), & x \le L\,; \\ v_x/(v_x + \gamma_\tau + \gamma_{\tau-L}), & x > L\,. \end{cases}$$

In this scenario, $x \to x + 1$, and

$$w_1(x + 1|\tau + 1) = P_1(x|\tau)\, w(x|\tau)\,. \tag{1}$$

No peptides are generated;

(2) Elementary event (b), which may not be followed by anything but the strand shift by one AA (as there is nothing to be cleaved). This scenario probability is

$$P_2(x|\tau) = \begin{cases} \gamma_\tau/(v_x+\gamma_\tau), & x \le L\,; \\ \gamma_\tau/(v_x+\gamma_\tau+\gamma_{\tau-L}), & x > L\,. \end{cases}$$

In this scenario, $x \to 1$, and

$$w_2(x|\tau+1) = \delta_{x,1} \sum_{x'=1}^{\infty} P_2(x'|\tau)\, w(x'|\tau)\,. \tag{2}$$

The peptides cut out are

$$Q_2(\tau, \tau-x+1|\tau) = P_2(x|\tau)\, w(x|\tau)\,; \tag{3}$$

(3) Elementary event (c), which may be followed either by strand shift (1) or by scenario (2). The probability of the first stage (c) is

$$P_c(x|\tau) = \begin{cases} 0, & x \le L\,; \\ \gamma_{\tau-L}/(v_x+\gamma_\tau+\gamma_{\tau-L}), & x > L\,. \end{cases}$$

After event (c), when $x \to L$, the number of the system states generated is

$$w_c(x|\tau) = \delta_{x,L} \sum_{x'=L+1}^{\infty} P_c(x'|\tau)\, w(x'|\tau)\,,$$

and the peptides cut out are

$$Q_c(\tau-L, \tau-x+1|\tau) = P_c(x|\tau)\, w(x|\tau)\,.$$

The subsequent events (1) or (2) should be considered as the respective above mentioned scenarios starting with the distribution $w_c(x|\tau)$, *i.e.*,

$$w_{c1}(x|\tau+1) = P_1(L|\tau)\, w_c(x-1|\tau) = P_1(L|\tau)\, \delta_{x,L+1} \sum_{x'=L+1}^{\infty} P_c(x'|\tau)\, w(x'|\tau)\,, \tag{4}$$

$$Q_{c1}(\tau-L, \tau-x+1|\tau) = P_1(L|\tau)\, Q_c(\tau-L, \tau-x+1|\tau) = P_1(L|\tau)\, P_c(x|\tau)\, w(x|\tau)\,; \tag{5}$$

$$w_{c2}(x|\tau+1) = \delta_{x,1} \sum_{x'=1}^{\infty} P_2(x'|\tau)\, w_c(x'|\tau) = \delta_{x,1}\, P_2(L|\tau) \sum_{x'=L+1}^{\infty} P_c(x'|\tau)\, w(x'|\tau)\,, \tag{6}$$

$$Q_{c2}(\tau-L, \tau-x+1|\tau) = P_2(L|\tau)\, Q_c(\tau-L, \tau-x+1|\tau) = P_2(L|\tau)\, P_c(x|\tau)\, w(x|\tau)\,, \tag{7}$$

$$Q_{c2}(\tau, \tau-x+1|\tau) = P_2(x|\tau)\, w_c(x|\tau) = \delta_{x,L}\, P_2(L|\tau) \sum_{x'=L+1}^{\infty} P_c(x'|\tau)\, w(x'|\tau)\,. \tag{8}$$

Collecting equations (1), (2), (4), (6), one finds the *master equation*

$$w(1|\tau+1) = \sum_{x=1}^{L} \frac{\gamma_\tau\, w(x|\tau)}{v_x+\gamma_\tau} + \left(1+\frac{\gamma_{\tau-L}}{v_L+\gamma_\tau}\right) \sum_{x=L+1}^{\infty} \frac{\gamma_\tau\, w(x|\tau)}{v_x+\gamma_\tau+\gamma_{\tau-L}}; \tag{9}$$

$$w(L+1|\tau+1) = \frac{v_L}{v_L+\gamma_\tau}\Big[w(L|\tau) + \sum_{x=L+1}^{\infty} \frac{\gamma_{\tau-L}\, w(x|\tau)}{v_x+\gamma_\tau+\gamma_{\tau-L}}\Big]; \tag{10}$$

$$w(x|\tau+1) = \frac{v_{x-1}\, w(x-1|\tau)}{v_{x-1}+\gamma_\tau+\Theta(x-L-1)\gamma_{\tau-L}} \quad \text{for } x \ne 1,\ x \ne L+1. \tag{11}$$

Here $x = 1, 2, 3, ..., M$ and $\tau = 1, 2, 3, ..., M-1$, where M is the length of the protein, and the Heaviside function $\Theta(x < 0) = 0$, $\Theta(x \geq 0) = 1$. Equations (9)–(11) form a linear map

$$w(x|\tau+1) = \sum_{y=1}^{\infty} \mathcal{L}_{xy}(\tau)\, w(y|\tau). \tag{12}$$

The whole contribution to the cleavage pattern

$$\begin{aligned} Q(\tau, \tau-x+1|\tau) &= Q_2(\tau, \tau-x+1|\tau) + Q_{c2}(\tau, \tau-x+1|\tau) \\ &= \frac{\gamma_\tau\, w(x|\tau)}{v_x + \gamma_\tau + \Theta(x-L-1)\gamma_{\tau-L}} + \frac{\delta_{x,L}\,\gamma_\tau}{v_L+\gamma_\tau} \sum_{x'=L+1}^{M} \frac{\gamma_{\tau-L}\, w(x'|\tau)}{v_{x'}+\gamma_\tau+\gamma_{\tau-L}}\,; \end{aligned} \tag{13}$$

$$\begin{aligned} Q(\tau-L, \tau-L-x+1|\tau) &= Q_{c1}(\tau-L, \tau-L-x+1|\tau) + Q_{c2}(\tau-L, \tau-L-x+1|\tau) \\ &= \frac{\gamma_{\tau-L}\, w(L+x|\tau)}{v_{L+x}+\gamma_\tau+\gamma_{\tau-L}}\,. \end{aligned} \tag{14}$$

All the rest [not specified by expressions (13), (14)] elements $Q(m, n|\tau)$ are zero. The expressions for digestion pattern $Q(m, n)$ after the processing of a single protein molecule are different for short polypeptides and long ones of a periodic AA sequence.

2. Short (25–50 AA) synthetic polypeptides

First we consider degradation of *short* (25–50 AA) *synthetic polypeptide* (protein), the most common situation for *in vitro* experiments. Here we start at $\tau = 1$ with $w(x|\tau = 1) = \delta_{x,1}$ and iterate linear map (12) till the last $\tau = M-1$. For a short polypeptide the releasing of the last fragment from the chamber at the "time moment" $\tau = M$ should be additionally taken into account: $Q(M, M-x+1|M) \to Q(M, M-x+1|M) + w(x|M)$. Hence, with $w(x|\tau)$ known for $\tau = 1, 2, ..., M$, one may evaluate digestion pattern $Q(m, n)$ from (13) and (14),

$$\begin{aligned} Q(\tau_1, \tau_2) &= Q(\tau_1, \tau_2|\tau_1) + \Theta(M-\tau_1-L)\, Q(\tau_1, \tau_2|\tau_1+L) \\ &= \delta_{\tau_1,M} w(\tau_1+L-\tau_2+1|M) + \frac{\gamma_{\tau_1}\, w(\tau_1-\tau_2+1|\tau_1)}{v_{\tau_1-\tau_2+1}+\gamma_{\tau_1}+\Theta(\tau_1-\tau_2-L)\gamma_{\tau_1-L}} \\ &\quad + \frac{\delta_{\tau_1-\tau_2+1,L}\,\gamma_{\tau_1}}{v_L+\gamma_{\tau_1}} \sum_{x=L+1}^{M} \frac{\gamma_{\tau_1-L}\, w(x|\tau_1)}{v_x+\gamma_{\tau_1}+\gamma_{\tau_1-L}} \\ &\quad + \Theta(M-\tau_1-L)\frac{\gamma_{\tau_1}\, w(\tau_1+L-\tau_2+1|\tau_1+L)}{v_{\tau_1+L-\tau_2+1}+\gamma_{\tau_1+L}+\gamma_{\tau_1}}, \end{aligned} \tag{15}$$

here $1 \leq \tau_2 \leq \tau_1 \leq M$. Since the protein may be cleaved starting both from the C- and from the N-terminal, the final digestion pattern is given by

$$Q_{\text{fin}}(\tau_1, \tau_2) = P_{\text{N}}\, Q_{\text{N}}(\tau_1, \tau_2) + P_{\text{C}}\, Q_{\text{C}}(M-\tau_2+1, M-\tau_1+1)\,. \tag{16}$$

The subscripts indicate which terminal goes first, P_{N} and $P_{\text{C}} = 1 - P_{\text{N}}$ are the probabilities of the degradation starting from the corresponding end. Generally, $v_{\text{N}}(x)$ and $v_{\text{C}}(x)$ may be slightly different, but here we neglect this difference. Note that

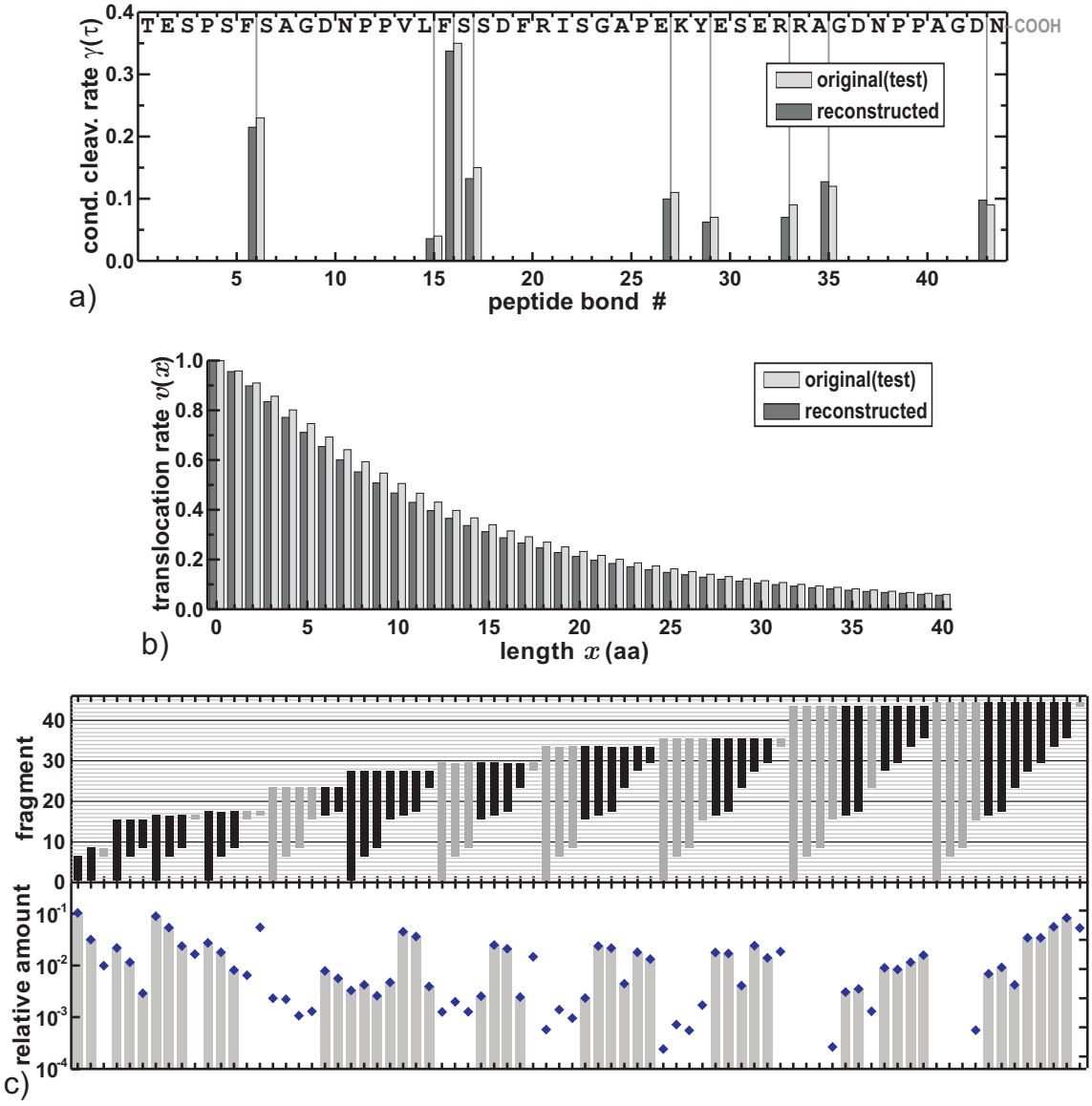

Figure 2. Test — Reconstruction of translocation rate function $v(x)$ and conditional cleavage rates $\gamma(\tau)$ for the 44mer peptide Kloe 316 [14, 15] [but with roughly estimated authentic (original) values of $\gamma(\tau)$], which is the subsequence 543–586 AA of human myelin associated glycoprotein. a) the conditional cleavage rates and the AA sequence; b) the translocation rate function; c) the upper plot presents the set of digestion fragments (black bars: fragments utilized for the reconstruction, grey bars: not utilized), and the lower plot presents the amount of the corresponding fragment (diamonds: the reconstructed values $Q_{\rm fin}$, grey bars: the values of $\widetilde{Q}$ utilized for the reconstruction).

a fragment length distribution $S(x)$ (often used in the literature [13]) is then the convolution

$$S(x) = \sum_{\tau=x}^{M} Q(\tau, \tau - x + 1)\,. \tag{17}$$

Digestion pattern $Q_{\rm fin}(\tau_1, \tau_2)$ is a functional of TRF $v(x)$ and CCR $\gamma(\tau)$. Utilizing MS data on the digestion pattern, one can determine nonzero values of $\gamma(\tau)$ (*i.e.*

positions of possible cleavage) and minimize the mismatch between $Q_{\rm fin}(\tau_1,\tau_2)$ and MS data $\widetilde{Q}(\tau_1,\tau_2)$ over $v(x)$, the nonzero values of $\gamma(\tau)$, and $P_{\rm N}$ in order to *reconstruct* them. Expecting the function $v(x)$ to be smooth, we parameterize appropriate approximate functions as

$$v_{\rm app}(x) = v_0 {\rm e}^{-\frac{A_2^2}{\sqrt{A_1^2+x}}+\frac{A_2^2}{|A_1|}-A_3^2(\sqrt{A_1^2+x}-|A_1|)}. \qquad (18)$$

Note, $v(x)$ and $\gamma(\tau)$ are defined up to a constant multiplier, which should be determined from the degradation rate in real time, but not from the digestion pattern.

In order to verify the robustness of the reconstruction procedure, numerous tests have been performed. A typical test presented in figure 2 has been performed in 4 steps:
(1) For given $v(x)$ [not generic for $v_{\rm app}$, *i.e.*, the used function $v(x)$ cannot be perfectly fitted with expression (18)] and $\gamma(\tau)$ digestion pattern $Q(\tau_1,\tau_2)$ has been evaluated.
(2) The result has been perturbed by the noise, $\widetilde{Q}_{\tau_1\tau_2} = Q_{\tau_1\tau_2} + 10^{-4} R_{\tau_1,\tau_2}\sqrt{Q_{\tau_1,\tau_2}}$, where R_{τ_1,τ_2} are independent random numbers uniformly distributed in $[-1,1]$.
(3) We have omitted the information about fragments, which relative amount is less than $5\cdot 10^{-3}$, and 1mer and 2mer fragments as being hardly detectable in experiments (one cannot distinguish identical AAs cut out from different parts of the polypeptide [16]).
(4) Resulting $\widetilde{Q}_{\tau_1\tau_2}$ has been used for the reconstruction of $v(x)$ and $\gamma(\tau)$.
The original and reconstructed data for $\gamma(\tau)$ (figure 2a) and $v(x)$ (figure 2b) are in a very good agreement. The reconstructed $P_{\rm N} = 0.52$ against original $P_{\rm N} = 0.50$.

Unfortunately, the data available in the literature are mainly too much incomplete (a lot of fragments are not accounted) and not enough precise for a truly reliable reconstruction [9] (the initial solutions used for experiments quite frequently contain not only the polypeptide to be digested but also a certain amount of its fragments, the first measurement of the proportions of the solution is performed to late, when considerably more than 5% of the initial substrate has been degraded and one may not neglect reentries of the digestion fragments into the proteasome, *etc.*).

Thus, we should note the limitations of the suggested reconstruction method:
• The reconstruction procedure for short polypeptides is very sensitive to measurement inaccuracy.
• For some polypeptides the procedure fails. This may happen due to a specific arrangement of cleavage positions, when different TRFs $v(x)$ provides almost identical digestion patterns.
• Though the whole information on $Q(\tau_1,\tau_2)$ is not needed, the number of nonzero values of $Q(\tau_1,\tau_2)$ required for a reliable (tolerant to noise) reconstruction is at least the twice number of reconstructed parameters, *i.e.* $2\times([\text{number of positions of potential cleavage}]+[\text{number of parameters of } v_{\rm app}]+1)$. For instance, for Kloe 258 in [9] the number of trustworthy and utilized values of $\widetilde{Q}(\tau_1,\tau_2)$ is 19 instead of the required $2\times(10+3+1) = 28$, it is a bit greater than the number of the unknown parameters, *i.e.*, 14. Hence, more accurate and comprehensive MS data on the digestion pattern are required.
• For short polypeptides the finishing stage of the degradation is relatively important, while in this stage the translocation rate is affected by the edge effects (the backward

end of the polypeptide gets inside the proteasome chamber) and is not the same as for the remainder of the polypeptide.

3. Long synthetic polypeptides of a periodic amino acid sequence

While a more comprehensive acquisition of data on digestion fragments and enhancement of experimental techniques for short polypeptides are up to experimentalists we propose experimental set-up which allows overcoming all the limitations mentioned above and is expected to be realizable. For this a *long synthetic polypeptide with a T-periodic AA sequence*: $\gamma(\tau) = \gamma(\tau + T)$ should be digested. Here "long" means one may neglect the peculiarities of the starting and finishing stages of the degradation, and $M \gg T$.

For the given direction of the degradation, *e.g.*, starting with the N-terminal, we are looking for the establishing T-periodic in τ solution $w_{\mathrm{N},T}(x|\tau) = w_{\mathrm{N},T}(x|\tau - T)$ to equation (12). The fragment (n, m) is identical to the one $(n+kT, m+kT)$, where k is integer; therefore $Q_{\mathrm{N}}(m, n)$ may be chosen to make contribution to $Q_{\mathrm{N}}(m - n + (n \bmod T), n \bmod T)$. The amount of fragments grows almost linearly with "time" τ as the polypeptide being processed. Hence, for the digestion pattern one finds

$$\begin{aligned} Q_{\mathrm{N},T}(\tau_1, \tau_2) &\equiv \lim_{\tau\to\infty} \frac{1}{\tau} \sum_{\tau'=1}^{\tau} Q_{\mathrm{N}}(\tau_1, \tau_2|\tau') = \frac{1}{T} \sum_{\tau'=1}^{T} Q_{\mathrm{N},T}(\tau_1, \tau_2|\tau') \\ &= \frac{1}{T}\Bigg[\frac{\gamma_{\tau_1}\, w_{\mathrm{N},T}(\tau_1 - \tau_2 + 1|\tau_1)}{v_{\tau_1-\tau_2+1} + \gamma_{\tau_1} + \Theta(\tau_1-\tau_2-L)\gamma_{\tau_1-L}} \\ &+ \frac{\delta_{\tau_1-\tau_2+1,L}\,\gamma_{\tau_1}}{v_L + \gamma_{\tau_1}} \sum_{x=L+1}^{\infty} \frac{\gamma_{\tau_1-L}\, w_{\mathrm{N},T}(x|\tau_1)}{v_x + \gamma_{\tau_1} + \gamma_{\tau_1-L}} + \frac{\gamma_{\tau_1}\, w_{\mathrm{N},T}(\tau_1 + L - \tau_2 + 1|\tau_1 + L)}{v_{\tau_1+L-\tau_2+1} + \gamma_{\tau_1+L} + \gamma_{\tau_1}}\Bigg] \end{aligned} \tag{19}$$

(here $1 \le \tau_2 \le T$ and $\tau_1 \ge \tau_2$).

To treat the degradation process starting with the C-terminal, one has (i) to perform the transformation $\gamma(\tau) \to \gamma(T - \tau)$, (ii) iterate linear map (12) with the new $\gamma(\tau)$ like for the N-case, but assuming $Q_{\mathrm{C}}(m, n|\tau)$ to make contribution to $Q_{\mathrm{C}}(m \bmod T,\, n - m + (m \bmod T))$. Unlike (16), the final result is

$$Q_{\mathrm{fin}}(\tau_1, \tau_2) = P_{\mathrm{N}} Q_{\mathrm{N},T}(\tau_1, \tau_2) + P_{\mathrm{C}} Q_{\mathrm{C},T}(T-\tau_2, T-\tau_1).$$

Matching $Q_{\mathrm{fin}}(m, n)$ to the MS data one can reconstruct $v(\tau)$, $\gamma(\tau)$, and P_{N}. For a test we have made use of the cleavage map of the digestion of yeast enolise-1 by human erythrocyte proteasome [17]. Looking at its subsequence 331–348 AA

...|ATAIEKKA|AD|ALLL|KV|NQ|...–COOH

(vertical stripes mark the positions of experimentally observed cleavages), one may expect the case, where the underlined subsequence is followed not by KV, but by KKA..., and the periodic sequence is

AD|ALLL|KKA| . . . |AD|ALLL|KKA| . . . |AD|ALLL|KKA–COOH ,

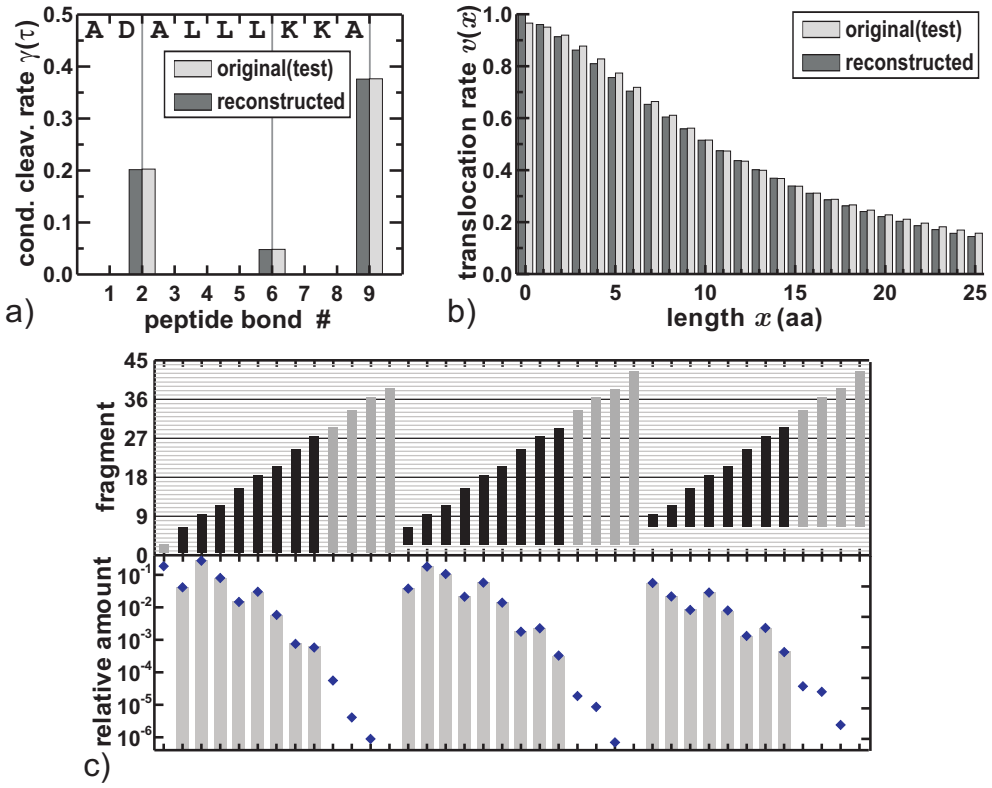

Figure 3. Test — Reconstruction of translocation rate function $v(x)$ and conditional cleavage rates $\gamma(\tau)$ for a 9-periodic polypeptide with the cleavage positions 2, 6, 9. (For description see caption to figure 2.)

to be realizable. For such a sequence a test like the one in figure 2 (but with much stronger dithering: $\widetilde{Q}_{\tau_1\tau_2} = Q_{\tau_1\tau_2} + 2\cdot 10^{-3} R_{\tau_1,\tau_2}\sqrt{Q_{\tau_1,\tau_2}}$) is presented in figure 3. Due to the small number of unknown parameters the reconstruction procedure is rather tolerant to measurement inaccuracy and does not require information on a large number of digestion fragments (the most easily detectable fragments are enough).

4. Long natural proteins

The case of a most immediate interest is the digestion of long natural proteins (over about 300 AA) because it concerns the *in vivo* proteasomal activity. A direct implementation of the procedure developed for short polypeptides is hardly possible here, as in the course of matching $Q(\tau_1, \tau_2)$ to the MS data, one has to perform a minimization over an enormous number of parameters. However, for long non-periodic proteins, one may assume $\gamma(\tau)$ to be a random process in order to evaluate some observable statistical properties like the fragment length distribution (FLD) of the digestion products, *i.e.* $S(x)$ [see equation (17)].

For this random process we adopt the following:

• the neighbor values $\gamma(\tau)$ and $\gamma(\tau+1)$ are mutually independent (what does not

necessarily mean that CCR $\gamma(\tau)$ is independent of neighbor AAs);
• $\gamma(\tau)$ is zero with a certain probability q, and has a finite probability density $g(\gamma)$ otherwise.

The normalized mean FLD $\mathcal{S}(x) \equiv \langle S(x)\rangle / \sum_{x'=1}^{\infty}\langle S(x')\rangle$ may be evaluated either via the plain iterating of (12)–(16) with noise $\gamma(\tau)$ over a large interval of τ or via the direct simulation of the system with a Gillespie algorithm (*e.g.*, see [18]). However, the calculation procedure may be considerably facilitated. For this purpose, let us average (12) over realizations of $\gamma(\tau)$,

$$\langle w(x|\tau+1)\rangle_\gamma = \textstyle\sum_{y=1}^{\infty}\langle \mathcal{L}_{xy}(\tau)\, w(y|\tau)\rangle_\gamma\,. \tag{20}$$

Noteworthy, $w(x|\tau)$ depends on $\gamma(\tau-1)$ and the preceding values of γ but is independent of $\gamma(\tau)$. Moreover, the impact of preceding values of γ decays in the course of the processing of the protein, and one may neglect the correlation between $w(x|\tau)$ and $\gamma(\tau-L)$ which are mutually distant in τ. Thus, $w(x|\tau)$ is independent of $\gamma(\tau)$ and $\gamma(\tau-L)$, which are involved in $\mathcal{L}_{xy}(\tau)$, and (20) yields

$$\langle w(x|\tau+1)\rangle_\gamma \approx \textstyle\sum_{y=1}^{\infty}\langle \mathcal{L}_{xy}(\tau)\rangle_{\gamma_\tau\gamma_{\tau-L}}\langle w(y|\tau)\rangle_\gamma\,; \tag{21}$$

from (13), (14), (17),

$$\begin{aligned}\langle S(x|\tau+1)\rangle_\gamma = \langle S(x|\tau)\rangle_\gamma &+ \left\langle \frac{\gamma_\tau}{v_x+\gamma_\tau+\Theta(x-L-1)\gamma_{\tau-L}}\right\rangle_{\gamma_\tau\gamma_{\tau-L}}\langle w(x|\tau)\rangle_\gamma \\ &+ \left\langle \frac{\gamma_{\tau-L}}{v_{L+x}+\gamma_\tau+\gamma_{\tau-L}}\right\rangle_{\gamma_\tau\gamma_{\tau-L}}\langle w(L+x|\tau)\rangle_\gamma \\ &+ \delta_{x,L}\sum_{x'=L+1}^{\infty}\left\langle \frac{\gamma_\tau}{v_L+\gamma_\tau}\cdot\frac{\gamma_{\tau-L}}{v_{x'}+\gamma_\tau+\gamma_{\tau-L}}\right\rangle_{\gamma_\tau\gamma_{\tau-L}}\langle w(x'|\tau)\rangle_\gamma\,,\end{aligned} \tag{22}$$

where

$$\begin{aligned}\langle f(\gamma_1,\gamma_2)\rangle_{\gamma_1\gamma_2} \equiv q^2 f(0,0) &+ q(1-q)\textstyle\int_0^\infty g(\gamma)[f(0,\gamma)+f(\gamma,0)]\mathrm{d}\gamma \\ &+ (1-q)^2\textstyle\int_0^\infty \mathrm{d}\gamma_1\int_0^\infty \mathrm{d}\gamma_2\, g(\gamma_1)\, g(\gamma_2)\, f(\gamma_1,\gamma_2)\,.\end{aligned}$$

The FLD observed in experiments is $\mathcal{S}(x)$ corresponding to the establishing steady solution $\langle w(x|\infty)\rangle$ to linear map (21).

Noteworthy, with the additional approximation

$$\langle \mathcal{L}_{xy}(\gamma_\tau,\gamma_{\tau-L})\rangle_{\gamma_\tau\gamma_{\tau-L}} \approx \mathcal{L}_{xy}(\langle\gamma\rangle,\langle\gamma\rangle)\,,$$

one may obtain an implicit recursive formula for establishing $\langle w(x|\tau)\rangle$ from (21),

$$\langle w(x+1|\infty)\rangle = \frac{(1+\delta_{x,L})\, v_x}{v_x+(1+\Theta(x-L))\langle\gamma\rangle}\langle w(x|\infty)\rangle\,, \tag{23}$$

and find FLD $\mathcal{S}(x)$ from (22),

$$\mathcal{S}(x) = \frac{\dfrac{(1+\delta_{x,L})\langle\gamma\rangle\ \langle w(x|\infty)\rangle}{v_x+(1+\Theta(x-L))\langle\gamma\rangle} + \dfrac{\langle\gamma\rangle\ \langle w(L+x|\infty)\rangle}{v_{L+x}+2\langle\gamma\rangle}}{\langle w(1|\infty)\rangle + \dfrac{\langle w(L+1|\infty)\rangle}{2}}\,. \tag{24}$$

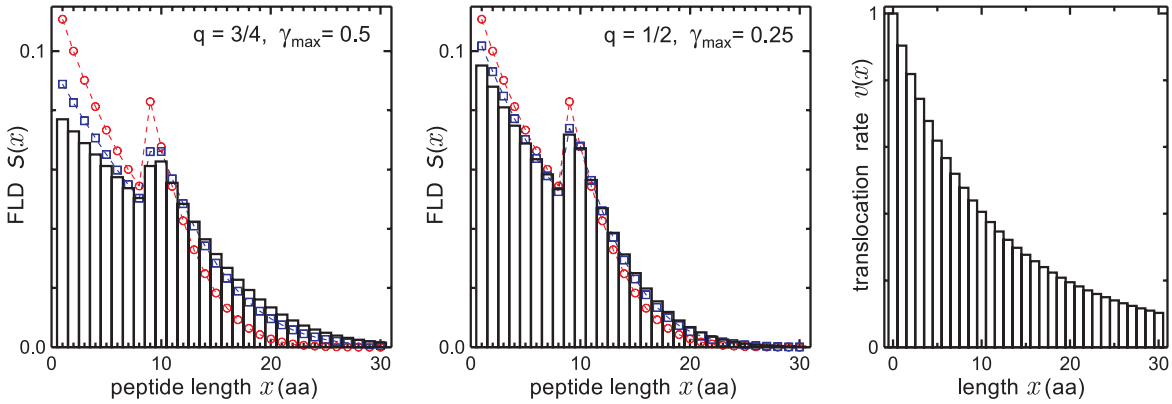

Figure 4. Samples of fragment length distribution $\mathcal{S}(x)$ (FLD) for the degradation of a long natural protein under the assumption, that conditional cleavage rate $\gamma(\tau)$ may be treated as a random process. The fraction q of nonscissile peptide bonds is indicated in the plots, nonzero values of $\gamma(\tau)$ are uniformly distributed in $[0, \gamma_{\max}]$, $L = 9$, the adopted translocation rate function $v(x)$ is plotted in the right plot. In two left plots, bars: results of the direct simulation with a Gillespie algorithm, squares: the approximation (21), (22), circles: the approximation (23), (24) with $\langle\gamma\rangle = (1-q)\gamma_{\max}/2$.

Remarkably, in the quasi-continuous limit (which is valid when $v(x)$ is a "slow" function of x), the last expressions provide (cf. [18])

$$\langle w(x|\infty)\rangle = (1+\Theta(x-L))\langle w(0|\infty)\rangle\, \mathrm{e}^{-\int\limits_0^x \frac{(1+\Theta(x-L))\,\langle\gamma\rangle}{v(x')}\mathrm{d}x'},$$

$$\mathcal{S}(x) = \langle\gamma\rangle\, \frac{\dfrac{\mathrm{e}^{-\int\limits_0^x \frac{(1+\Theta(x-L))\,\langle\gamma\rangle}{v(x')}\mathrm{d}x'}}{v(x)} + \dfrac{\mathrm{e}^{-\int\limits_0^{L+x} \frac{(1+\Theta(x-L))\,\langle\gamma\rangle}{v(x')}\mathrm{d}x'}}{v(L+x)}}{1+\mathrm{e}^{-\int\limits_0^L \frac{\langle\gamma\rangle}{v(x')}\mathrm{d}x'}}.$$

In figure 4, one may see, that the both above mentioned approximations become more accurate as q decreases. However, for realistic value $q \approx 3/4$ which is suggested by experimental cleavage maps (see figure 2, where the sites of a potential cleavage are taken from experimental data), the approximation (21), (22) works considerably better than the one (23), (24). Remarkably, as q increases with $\langle\gamma\rangle$ kept fixed, the local maximum near $x = L$ shifts from $x = L$ to higher values of fragment length x and the cutting-out of longer peptides becomes more probable. The existence of this maximum at $L \approx 8-10\,\mathrm{AA}$ deserves especial attention because the epitopes, involved in the functioning of the immune system and bound to MHC I molecules, have exactly such length [19].

The important limitation of this method is related to the reconstruction of $v(x)$ for 1mer and 2mer peptides. These peptides are hardly detectable in experiments and, therefore, experimental $\mathcal{S}(x)$ is not determined for $x = 1, 2$, and one cannot reconstruct the respective values of $v(x)$. Note, for methods suggested in sections 2 and 3

this limitation does not occur because, *e.g.*, for the subsequence `|F|S|SDFRISGAPE|` in figure 2, the information on $v(1)$ is reflected in the difference between the readily measurable amounts of generated peptides `|S|SDF...|` and `|SDF...|`, while for long natural proteins we lose the individual information on each specific peptide cut out.

5. Conclusion

In this paper we have discussed a model of the degradation of proteins by the proteasome which allows one to *reconstruct* the proteasomal translocation function and the cleavage specificity inherent to the amino acid sequence and not affected by proteasomal transport properties. With these properties determined, one can comprehensively predict digestion patterns of new proteins. The model is relevant for a broad variety of hypothetically possible translocation mechanisms [8, 11]. We have mathematically elaborated this model for the cases of (i) relatively short (25–50mers) synthetic polypeptides as the most common case for *in vitro* experiments, (ii) long periodic polypeptides (proposed experiments with such polypeptides are very promising for reverse engineering), and (iii) long natural proteins.

In [18], we have already discussed how peculiarities of the translocation function may lead to the multimodality of the fragment length distribution even for $\gamma(\tau) = const$. Here we have shown that the amount of each digestion fragment is not only determined by the cleavage map [specifically, conditional cleavage rate $\gamma(\tau)$] of the substrate but is also crucially affected by nonuniformity of the translocation rate. The results of implementation of the developed theory for processing experimental data on digestion patterns for different proteasome species under different conditions can give insight into the nature of the protein translocation mechanism inside the proteasome. They can as well elucidate the unanswered question whether there is some preference for starting the degradation with the N- or C-terminal of the protein, and how this preference is affected by regulatory complexes. Hopefully, theoretical results will stimulate new experiments as suggested in this paper for the case of a periodic polypeptide.

Acknowledgments

We are thankful to Susanne Witt and Michele Mishto for useful discussions. The work has been supported by grants of the VW-Stiftung, the BRHE program, and the Foundation “Perm Hydrodynamics”.

References

[1] Rock K L, Gramm C, Rothstein L, Clark K, Stein R, Dick L, Hwang D and Goldberg A L 1994 *Cell* **78** 761–71

[2] Kloetzel P M 2001 *Nat. Rev. Mol. Cell. Biol.* **2** 179–88

[3] Tanahashi N, Murakami Y, Minami Y, Shimbara N, Hendil K B and Tanaka K 2000 *J. Biol. Chem.* **275** 14336–45

[4] Rubinsztein D C 2006 *Nature* **443** 780–86
Mishto M, Bellavista E, Santoro A and Franceschi C 2007 *Central Nervous System Agents in Medicinal Chemistry* **7** 236–40
[5] Holzhütter H G and Kloetzel P M 2000 *Biophys. J.* **79** 1196–205
[6] Peters B, Janek K, Kuckelkorn U and Holzhütter H G 2002 *J. Mol. Biol.* **318** 847–62
[7] Luciani F, Kesmir C, Mishto M, Or-Guil M and de Boer R J 2005 *Biophys. J.* **88** 2422–32
[8] Zaikin A and Pöschel T 2005 *Europhys. Lett.* **69** 725–31
[9] Goldobin D S, Mishto M, Textoris-Taube K, Kloetzel P M and Zaikin A 2008 Reverse engineering of proteasomal translocation rates [*submitted*, preview: arXiv:0804.0682]
[10] Witt S (private communication)
[11] Reimann P, Van den Broeck C, Linke H, Hänggi P, Rubi J M and Perez-Madrid A, 2001 *Phys. Rev. Lett.* **87** 010602
[12] Tenzer S, Peters B, Bulik S, Schoor O, Lemmel C, Schatz M M, Kloetzel P M, Rammensee H G, Schild H and Holzhütter H G 2005 *Cell. Mol. Life Sci.* **62** 1025–37
[13] Kisselev A F, Akopian T N and Goldberg A L 1998 *J. Biol. Chem.* **273** 1982–9
[14] Mishto M *et al* 2006 *Biol. Chem.* **387** 417–29
[15] Mishto M, Luciani F, Holzhütter H G, Bellavista E, Santoro A, Textoris-Taube K, Franceschi C, Kloetzel P M and Zaikin A 2008 *J. Mol. Biol.* **377** 1607–17
[16] Kohler A, Cascio P, Leggett D S, Woo K M, Goldberg A L and Finley D 2001 *Mol. Cell* **7** 1143–52
[17] Nussbaum A K, Kuttler C, Hadeler K-P, Rammensee H-G and Schild H 2001 *Immunogenetics* **53** 87–94
[18] Zaikin A, Mitra A K, Goldobin D S and Kurths J 2006 *Biophys. Rev. Lett.* **1** 375–86
[19] Falk K, Rötzschke O, Stevanovic S, Jung G and Rammensee H G 1991 *Nature* **351** 290–6
Madden D R, Gorga J C, Strominger J L and Wiley D C 1991 *Nature* **353** 321–5