University of Groningen

(Membrane) Protein Production in Context

Schavemaker, Paul E.; Poolman, Bert

Published in:
Trends in Biochemical Sciences

DOI:
[10.1016/j.tibs.2018.08.009](https://doi.org/10.1016/j.tibs.2018.08.009)

IMPORTANT NOTE: You are advised to consult the publisher's version (publisher's PDF) if you wish to cite from it. Please check the document version below.

Document Version
Publisher's PDF, also known as Version of record

Publication date:
2018

[Link to publication in University of Groningen/UMCG research database](#)

Citation for published version (APA):
Schavemaker, P. E., & Poolman, B. (2018). (Membrane) Protein Production in Context. *Trends in Biochemical Sciences*, 43(11), 858-868. <https://doi.org/10.1016/j.tibs.2018.08.009>

Copyright

Other than for strictly personal use, it is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license (like Creative Commons).

The publication may also be distributed here under the terms of Article 25fa of the Dutch Copyright Act, indicated by the "Taverne" license. More information can be found on the University of Groningen website: <https://www.rug.nl/library/open-access/self-archiving-pure/taverne-amendment>.

Take-down policy

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Downloaded from the University of Groningen/UMCG research database (Pure): <http://www.rug.nl/research/portal>. For technical reasons the number of authors shown on this cover page is limited to 10 maximum.

Opinion

(Membrane) Protein Production in Context

Paul E. Schavemaker¹ and Bert Poolman^{1,*}

Great progress has been made in elucidating the structural and mechanistic basis of (membrane) protein production. Here, we attempt to look ahead and indicate four directions in which our understanding of the protein production process can grow: (i) determine how the molecular mechanisms influence higher-level processes, such as the distribution of protein copy number over a population of cells or the cell growth rate; (ii) explore the functional landscape that the molecular mechanisms of protein production exist in, for instance by comparing membrane protein insertion mechanisms; (iii) uncover the life history of proteins – that is, what happens to them between their synthesis and degradation; and (iv) determine, and connect by calculation, the numbers that are associated with (membrane) protein production.

The Context of Molecular Mechanisms

Over the past decades great progress has been made in unraveling the molecular mechanisms underlying many processes in biology, including the production of (membrane) proteins. Atomic structures have been determined for the ribosomes and the Sec translocon as well as of many of the proteins that service these foci of (membrane) protein production [1–3]. Many of the interactions and mechanisms by which the constituents of the molecular machinery operate were also deciphered. We suggest that enough is known about the molecular mechanisms of protein production to put them in context. We describe four ways in which this can be done: (i) by connecting what we know of molecular mechanisms to more abstract (higher-level) phenomena such as the distribution of protein copy number over a population of cells, macromolecular crowding, spatial organization of protein synthesis, or the cell growth rate; (ii) by exploring the functional landscape that the production machinery operates in, for instance by comparing membrane protein insertion mechanisms; (iii) by examining the life history of proteins – that is, what happens to them between their synthesis and degradation; and (iv) by tying together protein production and other cellular processes by considering abundances, rates, etc.

Connecting the Levels of (Membrane) Protein Production

It is hard to imagine a cellular or organismal process that is not underlain in some way by the molecular mechanism of protein production. Yet it is unclear how, or even whether, the details of the molecular protein production machinery determine the nature of the higher-level phenomena. For example, do the properties of the ribosome determine what equations describe cell growth rate or the distribution of protein copy numbers over a population of cells? In this section we summarize the levels of (membrane) protein production. We start at the level of a population of cells and work our way down to the level of molecular mechanisms. Where possible we indicate how the levels are connected.

A single cell in solution will produce other cells. Although ultimately a consequence of molecular machinery such as the ribosome, at the level of whole cells the production process can be described simply by an exponential function that relates the initial cell number, N_0 , to later ones,

Highlights

The process of (membrane) protein production has been studied on many levels of abstraction, linking cell growth via protein copy number distributions to the molecular mechanisms of protein production.

The machineries of membrane protein production (Sec translocon, Tat system, TIM and TOM complexes, etc.) have similarities that can be exploited to study the functional landscape that these machineries are a part of.

Protein production machinery and its relation with chaperones and proteases define a fate map for proteins, which contains the life history of proteins.

Numbers, concentrations, rates, diffusion coefficients, etc. are known for many elements of the protein production process, with varying credibility.

¹Department of Biochemistry, University of Groningen, Groningen, The Netherlands

*Correspondence: b.poolman@rug.nl (B. Poolman).

N_t , and some parameter, k , to the rate of change (Figure 1A):

$$N_t = N_0 e^{kt}. \quad [1]$$

The shape of this function is not in any way dependent on the molecular mechanisms of protein production. It is simply a logical consequence of the fact that the rate of production of the material is linearly proportional to the amount of material there already is and thus is a fundamental property of the world. The value of the parameter k , by contrast, can depend on how fast ribosomes incorporate amino acids into polypeptide chains. Of course, at some stage the nutrients will run out and the rate of multiplication will decline. This can be described by the following equation [4], which has an 's' shape (Figure 1A):

$$N_t = \frac{KN_0 e^{kt}}{K + N_0(e^{kt} - 1)}. \quad [2]$$

Here K is the maximum number of cells – the carrying capacity. Again the shape of the function is not determined by the molecular mechanisms of protein production.

Trends in Biochemical Sciences

Figure 1. Illustration of Equations. (A) Exponential growth with and without carrying capacity. Carrying capacity, K , is 10 000, $N_0 = 1$, and $k = 1$ for both curves. The unit of time is arbitrary. (B) Variation of mRNA number, n , in a population of cells, assuming that the production of mRNA molecules can be described by a Poisson process. λ is the mean number of mRNAs per cell. (C) Variation of protein number, n , in a population of cells. Generated with the gamma distribution. The number of mRNAs made in a cell cycle, a , is varied and the values are indicated in the plots. The mean number of proteins made per mRNA, b , is ten.

Besides increasing in number, cells can change their character (i.e., proteome, metabolome) over time in response to some environmental challenge. Cells can become smaller [5] or filamentous [6] or they develop the ability to swim around [7]. Again some form of production process may be at work here. Fast changes of the character of a cell are not in the purview of protein production; for example, the response to an osmotic up- or downshift is initially performed by membrane proteins and lipids (e.g., water permeation) that are made in advance – at a later time, protein production can come into play. What kind of changes can happen to a cell probably depends mostly on what proteins and RNAs themselves can do and not on how they are made. An exception is maintaining long cells, which is more difficult in cases where the molecular machinery of protein production is localized to a particular spot (e.g., RNA polymerase localized to the nucleus in neurons).

Looking at a single cell, we do not only deal with particular proteins but also with the concentration of all proteins together. This bulk protein can have a possible impact on many processes in the cell due to the phenomenon of excluded volume (macromolecular crowding), which can change the equilibrium and rate constants of reactions and the diffusion coefficients of molecules [8–10]. Thus, we have a coupling between the phenomenon of protein production and many other processes in the cell that is independent of the production of the protein types involved directly in those processes. The molecular machinery of protein production sets most likely the total protein concentration, but it is not known how.

The copy numbers/concentrations of a single type of protein vary between cells in a population due to chance, and in *Escherichia coli* this can be described, in many cases, by a gamma distribution (Figure 1C) [11]:

$$P(n) = \frac{n^{a-1} \times e^{-\frac{n}{b}}}{b^a \times \Gamma(a)}. \quad [3]$$

Here $P(n)$ is the probability for a cell to have n proteins, a is the mean number of mRNAs produced per cell cycle, b is the mean number of proteins produced per mRNA, and $\Gamma(a)$ is the gamma function of a . The protein concentration is obtained by dividing the protein number by the product of the volume of the cell and Avogadro's number. The gamma distribution was tested in *E. coli* cells whose protein numbers were determined after labeling the proteins with a fluorescent tag [11]. For 1009 of the 1018 genes tested, the protein copy number distribution followed a gamma distribution. However, at larger mean copy numbers the relative variability of copy numbers over the population reached a lower limit at about 30% [11]. The derivation of the gamma distribution depends, among other things, on the fact that mRNA production events do not influence each other (Poisson process). This lack of influence is a molecular feature that could be different for different genes; for example, by having a positive feedback loop. This loop reinforces fluctuations in mRNA production causing a bimodal distribution rather than a gamma distribution.

So far we have treated individual cells as if they are uniform entities without 3D shape or size. However, we know of a number of examples where the production of proteins is not homogeneous over space: there is no protein production in the eukaryotic nucleus, the ribosomes in *E. coli* are excluded from the nucleoid under some conditions [12], some RNA degradation proteins in *E. coli* are localized to the membrane [13], genes in *E. coli* have specific locations in the cell [14], and different proteins that are part of the same complex can be produced from the same mRNA and thus start in close proximity [15]. In this last example there need not be spatial variation of the complex in the cell, but producing the proteins that constitute the complex close together could reduce misfolding and aggregation. To take into account spatial

variation, one has to include diffusion. This can be done analytically by using the diffusion equation, which relates the rate of change in concentration to the concentration profile over space [16]. A more detailed way to include spatial variation is to treat every mRNA, protein, ribosome, and degradation machinery as a diffusing dot that can undergo reactions [17]. As described above, the localization of various elements of the molecular machinery puts limits on what cells can do.

The production of mRNAs by RNA polymerase may work like a Poisson process, with an exponential distribution in the time between mRNA production events and with the number of mRNAs being produced in a fixed time window following a Poisson distribution (Figure 1B):

$$P(n) = \frac{e^{-\lambda} \lambda^n}{n!}. \quad [4]$$

Here $P(n)$ is the probability that n mRNAs are produced in a fixed window of time and λ is the mean number of mRNAs produced in this same time window. For this to hold, the number of (free) RNA polymerases needs to be constant over time, which gives the constant transcription initiation probability. However, in at least some cases mRNAs are produced in bursts; for example, due to the influence of DNA supercoiling on transcription [18]. If this is the case, the spread in the number of mRNAs produced in a fixed time window will be broader than the Poisson distribution predicts [19]. The mean number of mRNAs produced in a window of time is a consequence of the strength of the promoter and, in limiting cases, of the rate at which the RNA polymerase can synthesize a single mRNA. The small numbers of molecules involved causes the stochasticity of the production of mRNAs and the molecular mechanism of mRNA production is unable to prevent it.

Monitor a single mRNA and every once in a while you will witness the arrival of a ribosome. The ribosome will produce a single protein and leave again. If the lifetime of an mRNA is a single fixed time, the distribution of proteins produced from a single mRNA will also follow a Poisson distribution. However, it turns out that the lifetime of mRNAs, at least in *E. coli*, is exponentially distributed [20]. This leads to an exponential distribution also for the number of proteins produced per mRNA [19]. You expect the time between protein production events from the same mRNA to be distributed exponentially as well, because there is a constant number of (free) ribosomes. Differences in the localization of mRNA could influence how many proteins are produced from it. For example if an mRNA remained, by chance, in the nucleoid region it will encounter fewer ribosomes and be translated less often before finally being degraded. Additionally, different copies of the same type of mRNA could harbor a different complement of RNA-binding proteins, which could also affect the protein production rate. Studies in eukaryotes have shown that not all mRNAs produce proteins at the same rate [21]. In yeast, different mRNA variants of the same gene can have different lifetimes [22], which can affect the number of proteins produced per mRNA. These mechanisms of mRNA production/maintenance that are different in eukaryotes could lead to deviations from the gamma distribution shown above for protein copy numbers in *E. coli*.

Monitor a single ribosome and you will witness the coming into being of a single protein. You will see the ribosome assembling the protein chain one amino acid at a time [23]. Occasionally, when encountering a particular grouping of codons or mRNA structure, the ribosome will move more slowly [23,24]. These peculiarities in the formation of amino acid chains may average out so that they can be ignored when studying how fast full proteins are made. In other cases they may lead to lethal aggregates of protein or the formation of inclusion bodies. On what are probably rare occasions there is a problem with the mRNA and translation has to be aborted

with the tmRNA system [25]. It will take the ribosome a certain amount of time to finish the single protein. The distribution of these times is likely to be Gaussian. The reason for this is that the protein is made in hundreds of steps each with some distribution associated with it. According to the central limit theorem it does not matter what the distributions are for the steps, because, as long as you have many steps contributing to one final distribution, this distribution will be Gaussian (see also [19]). Thus, the fact that the distribution of protein production times is Gaussian is independent of the molecular mechanism of translation.

In the preceding we have in places indicated how the molecular mechanisms link to the higher-level processes. What we have shown here no doubt is only a smidgen of what is out there. We implore molecular biologists to not only determine the molecular mechanisms of production processes but also consider how each of these mechanisms allows, negates, or is irrelevant to higher-level processes.

The Functional Landscape of Proteins

The macromolecules that carry out the bulk of the function in any organism are all made of the same types of molecules, most of protein and some of RNA. The functions they perform are varying combinations of the basic phenomena of diffusion, conformational changes, binding, and chemical reactions. This means that all molecular mechanisms are functionally related and are thus part of what we here call the functional landscape. On this landscape we can trace a path between any two molecular phenomena or mechanisms. However, molecular biologists treat the molecular mechanisms of protein production as mostly unrelated to the rest of the possible molecular mechanisms. Connecting the molecular mechanisms of protein production to other mechanisms is a challenge, but we make a start by noting some similarities between them. A brief overview of the molecular machinery underlying (membrane) protein production is given in Box 1. The ribosome and DNA and RNA polymerases [26,27] are similar in that they are all template-based polymerases. The Sec translocase is similar to YidC [28], the TIM and TOM complexes [29], and the Tat translocase [30] in that it transports proteins into or over the membrane. The Sec translocase also shares similarity with DNA helicases [31], as they both let polymers thread through them. No doubt there will be many more similarities between seemingly distinct molecular processes and we suggest that it is important to map these

Box 1. Molecular Machinery Underlying Protein Production

Here we give a short overview of the different macromolecules that play a role in protein production. In a sense, all elements of a cell are part of the production process because the reason for a cell's existence is to make more cells. We focus on the machinery that performs the synthesis and membrane insertion of proteins in *Escherichia coli*, but these mechanisms are similar in other organisms. As soon as the ribosome-binding site is made, a ribosome can bind and start synthesizing a protein; this is before the mRNA has been fully synthesized. The ribosome-binding site on the mRNA is bound by a complex of the small ribosomal subunit (30S), formylmethionine-tRNA, and the initiation factors IF1 and IF3; the initiation factor IF2 helps to deliver formylmethionine-tRNA to the small ribosomal subunit. Next, the large ribosomal subunit (50S) binds [45]. Starting at the formylmethionine, amino acids are added to form an amino acid chain in cycles of conformational changes of the ribosome and assisted by the elongation factors EF-Tu and EF-G [46]. The specific amino acids come attached to specific tRNAs, to which they are coupled by specialized amino acyl tRNA synthetases [47]. When the amino acid chain is done, signaled by a stop codon, the release and recycling factors RF1, RF2, RF3, and RRF come in to release the nascent chain and dissociate the small and large ribosomal subunits [45].

For membrane proteins there is a part of the amino acid sequence that, as soon as it finds its way out of the ribosome, binds to a protein-RNA complex called the signal recognition particle (SRP). The SRP then guides the membrane protein-ribosome complex via the peripheral membrane protein FtsY to the Sec translocon. The Sec translocon, also called SecYEG, comprises the proteins SecY, SecE, and SecG and facilitates translocation of the membrane protein into the membrane. The part of the membrane protein that is first to emerge out of the ribosome can be inserted before the rest of the protein has been made [48].

connections. Related to this map we can also ask questions such as: why does the Sec translocon carry out transport both into and over the membrane? It may be that, if we were able to separate insertion into and transport across the membrane, we could produce membrane proteins more efficiently; that is, by reducing the constraints for optimization.

Placing molecular machineries in a functional landscape may help us to understand what makes them tick. For instance, we could, in this landscape, step away from the molecular machinery of translation by making a change in the ribosome or other components and ask whether this is still the same type of machine. If the change caused a difference in its spatial organization of protein production [32], in the regulation of protein production by ribosome dimerization [33], or in the fact that ribosomes bind to positively charged proteins [34], we would not change the type of machine that the ribosome is. A change that made the interaction between tRNA and ribosome impossible would change the type of machine that the ribosome is. Creating a functional landscape is not just for understanding the molecular mechanisms of protein production but also for finding new kinds of mechanisms, which may or may not be closely related to the protein production mechanism. In time we may be able create a description of protein production that makes these processes continuous with other now-existing, once-having-existed, and perhaps someday-to-be-created molecular mechanisms.

The Life History of Proteins

The life history of an organism is what happens to it during its lifetime. Similar to an organism, a protein is made, performs work, and then gets inactivated or degraded; it thus has a life history. Of key importance to the life history of proteins are the molecular machineries of protein maintenance: the chaperones and the proteases. Some of these are described in Box 2. If you put together everything that happens to a protein, and with what probability, you would end up with a protein fate map. In filling in this fate map, you would ask the following questions. What fraction of all amino acid chains whose synthesis is started (i.e., two amino acids coupled together) actually reach a full chain? What fraction of full amino acid chains fold properly? What fraction of folded proteins performs what number of actions (e.g., translocation of molecules across membranes) during its lifetime? With what frequency does an amino acid chain (completely synthesized or not) encounter a chaperone? What fraction of those interactions is consequential? With what frequency does a protein (in the folded or unfolded state) encounter each of the other proteins in the cell? With what frequency does a protein encounter the membrane? With what frequency does the protein unfold? How many cell cycles does a protein last? How many minutes does a protein last? What fraction of cell volume does the

Box 2. Molecular Machinery Underlying Protein Quality Control

In *Escherichia coli* cells, the processes of translation and insertion do not produce and insert proteins in isolation. There is a plethora of proteins (chaperones) that assures that proteins are correctly, and efficiently, folded. These are, for example, trigger factor, DnaJ, DnaK, GrpE, SurA, and the GroEL/ES protein complex [49]. Trigger factor is associated with ribosomes and interacts with most proteins when they are being synthesized, preventing hydrophobic patches from interacting. DnaJ, DnaK, and GrpE facilitate protein folding in an ATP-dependent manner. SurA isomerizes prolines in proteins so that the proteins can fold properly. Finally, the GroEL/ES is a big complex that traps proteins inside a big cavity and assists in their folding. The proteins are degraded by so-called proteases; for example, in response to a change in environment that needs a change in the set of proteins in the cell. In *E. coli* there are the ClpAP, ClpXP, Lon, and HslUV proteases. These specifically degrade a (broad) subset of proteins [50].

For membrane proteins there are also various factors that help protein production along. In *E. coli* we have SecD, SecE, YajC, YidC, YidD, and FtsH. The SecDF(YajC) complex assists in the insertion of membrane proteins. YidC assists SecYEG but can also insert some membrane proteins by itself [48]. YidD has been associated with the insertion of YidC-dependent substrates [51]. FtsH degrades misassembled membrane proteins [52].

protein sample per unit time? Some of these numbers are reported in Table 1. Answers to these questions help in understanding what would happen to a protein when some of the cellular parameters change; for example, if the concentration of the chaperone GroEL/ES changes or if the diffusion rate becomes slower after an osmotic upshift.

In considering the protein maintenance systems we can ask an important question: what functions of cells would not be possible without such systems? For example, are chaperones and proteases necessary for the existence of the Sec translocase? Here we do not mean what systems will be crippled when you knockout the chaperones or proteases in some organism. Because in principle the cell could be arranged in a different manner, the need for chaperoning would be avoided. For instance, you could adjust the amino acid sequence to make proteins better able to fold so they can do so without assistance, or you could use a molecular system that has the same effect but works by a different mechanism that is affected less by folding problems. We are asking whether there are functions that can never be performed if it were not

Table 1. Protein Production Numbers for *Escherichia coli*

Number of genes	4318, 4505	KEGG database (strain MG1655), [37] (strain MG1655)
Number of genes encoding proteins	4140	KEGG database (strain MG1655)
Number of genes encoding membrane proteins	~900	Helix bundle membrane proteins, [38]
Mean protein length	318 amino acids	Calculated from data from the KEGG database (strain MG1655), the mean is over genes and does not include protein copy numbers
Ribosome copy number per cell	6800–72 000, 55 000	Depends on cell growth rate [12,39]
RNA polymerase copy number per cell	1500–11 000, 4600	Depends on cell growth rate [12,39]
GroEL complex copy number per cell	180–7000	Depends on cell growth rate [39]
Number of proteins per μm^3 cell volume	2.7×10^6	[40]
Cell volume	0.4–2.5 μm^3	[5]
Frequency of encountering another protein	1 s ⁻¹	In <i>E. coli</i> , average-sized protein [41]
Protein chain elongation rate	10–20 amino acids/s	[42]
RNA chain elongation rate (mean)	25 nt/s	[20]
Average mRNA synthesis time	133 s	[20]
Average mRNA lifetime	4.1 min	Different from value mentioned in the paper but calculated from the same data, [20]
Protein lifetime	2 to >70 h	For 90% of proteome, during growth, some proteins are degraded within minutes [43]
Average mRNA length	1286 nt	[20]
mRNA diffusion coefficient	0.04 $\mu\text{m}^2/\text{s}$	When bound to ribosome, [12]
Ribosome diffusion coefficient	0.04 $\mu\text{m}^2/\text{s}$	[12]
SecY copy number	291	Calculated from PaxDb (integrated) assuming a total protein number of 3×10^6 , [44]
SecE copy number	105	See above
SecG copy number	549	See above
YidC copy number	210	See above
Ffh (SRP subunit) copy number	519	See above
FtsY copy number	645	See above
SecA copy number	765	See above

Box 3. Verifying Protein Production Numbers by Calculation

Protein production is automatically associated with questions of how many, and how fast? How many proteins are there in a single *Escherichia coli* cell? For a cell with a volume of $1\ \mu\text{m}^3$ there are about 2.7×10^6 proteins (see Table 1 in main text). All of these proteins need to be made by the complement of ribosomes present in the cell. There are about 30 000 ribosomes in a cell with a volume of $1\ \mu\text{m}^3$ [39]. Each cell cycle, these ribosomes thus have to produce ~ 100 proteins each. All this is with a cell division time of 40 min [39,53]. How fast can a ribosome produce 100 proteins? The mean protein length is about 300 amino acids (see Table 1 in main text), which is based on the number of genes that encode a protein of a certain length; it does not take into account protein abundance. The translation rate is 10–20 amino acids/s. This gives a translation time of 15–30 s per protein. Multiplying by the total number of proteins gives a total time of 25–50 min. This matches, albeit coarsely, with the cell division time. Each of the numbers mentioned here, whether measured or calculated, can be used to test hypotheses. However, there is more to the calculations. They validate the numbers by allowing cross-referencing.

for protein maintenance systems. Some answers may be found by looking at what protein maintenance systems and what functions are present in the simplest microbes.

Protein Production Numbers for *E. coli*

The numbers associated with cell processes, be they protein copy numbers, reaction rates, or others, are increasingly becoming known [35,36]. Here we bring together some of these numbers important to protein production in *E. coli* (Table 1). It is important not just to let these numbers stand on their own but to connect them by calculation, and thereby verify them by checking for contradictions. We do this for protein production in general in Box 3 and membrane protein production specifically in Box 4. These calculations show that for protein production the numbers show agreement, whereas for membrane protein production the numbers show a conflict. The calculation on membrane protein production suggests that the Sec translocon copy numbers determined in proteomics studies are too low. We hope that calculations like these will be used as a guide to what properties are measured, so that faulty measurements and assumptions can be rooted out.

For the examples worked out in Box 3 and 4 there is a heavy reliance on average numbers. This ignores the differences between cells in a population and the differences between genes in a single cell. Distributions have been determined for protein and mRNA length, mRNA lifetime,

Box 4. Verifying Membrane Protein Production Numbers by Calculation

Following their synthesis, membrane proteins have to be inserted into the membrane. This provides another constraint on the number of proteins that are produced per cell cycle, because this depends on the number of SecYEG molecules and the rate with which they translocate proteins. The PaxDb contains abundance data for *Escherichia coli* and shows that there are great differences between the measurements for a single protein. For example, the copy numbers for SecY range from nine to 1764. A dataset that integrates all of the data gives a copy number of 291 for SecY. This should be compared with 105 for SecE and 549 for SecG, both of which reside with SecY in the same complex. Numbers for these and other players in membrane protein production are listed in Table 1 in main text. We assume that translocon numbers are the same as SecY numbers, and thus about 300. On average, there are about three transmembrane helices per membrane protein ([54], data from *Acinetobacter baumannii*; may be an underestimate for *E. coli* [55]). Each transmembrane helix contains about 20 amino acids, and assuming that loops account for another 40 amino acids we end up with a total of 100 amino acids per membrane protein. Let us also assume that the insertion machinery works as fast as the ribosome: 10–20 amino acids/s (this gives an upper limit to the insertion rate). This means that it takes 5–10 s for a translocon to insert a single membrane protein. Each translocon can insert 240–480 proteins per 40-min cell cycle. The 300 translocons together can insert a total of 72 000–144 000 membrane proteins. This is probably an overestimate, as our number of transmembrane helices and loop residue number are on the low side. Taking the area per membrane protein to be $4.5\ \text{nm}^2$ [54] and the surface area of a cell (spherocylinder) of diameter $0.8\ \mu\text{m}$ and length $2.2\ \mu\text{m}$ [39], the membrane fraction covered with membrane protein is 6–12%. This is somewhat lower than estimates reported elsewhere for various types of membranes (>20%) [54]. Given that the translocon is also responsible for the transport of periplasmic, outer membrane, and extracellular proteins, the number of 300 copies per cell seems too low. So, unlike the calculation in Box 3, here we seem to have found a contradiction.

Trends in Biochemical Sciences

Figure 2. Genome-Wide Parameters for Protein Production in *Escherichia coli*. Parameters for proteins in red; parameters for mRNA in blue. (A) Protein length distribution. Taken from the KEGG database entry on *E. coli* MG1655. (B) mRNA lifetime distribution [20]. (C) mRNA length distribution [20]. (D) Transcription elongation rate distribution (for mRNA) [20]. (E) mRNA synthesis time distribution. Calculated from mRNA length and transcription elongation rate [20]. (F) Distribution of mRNA copy number per cell [11]. (G) Distribution of protein copy number per cell; about 1000 proteins are represented [11]. (H) Distribution of protein copy number per cell. Combination of datasets from PaxDb, with about 4000 proteins [44]. Note that the PaxDb numbers were multiplied by three because they were reported as parts per million and the number of *E. coli* proteins is about 3 million (Table 1).

transcription elongation rate, mRNA synthesis time, and mRNA and protein copy number per cell. Histograms are presented in Figure 2. There are interesting observations in here that cry out for further consideration. In the distributions of mRNA lifetime, transcription elongation rate, and mRNA synthesis time there are some considerable outliers. For example, there are a number of mRNAs that take more than 10 min to be synthesized. It is important to know how facts like this tie into the production and functioning of these proteins.

Concluding Remarks

Understanding of the molecular mechanisms of protein production is now fairly complete. We have here outlined four ways in which we can put these findings in context (see Outstanding Questions). (i) Connecting the molecular mechanisms to higher-level phenomena so that we can judge the importance of such mechanisms. We argue that in some cases molecular details are irrelevant for understanding higher-level phenomena. (ii) Comparing different types of molecular machinery involved in protein production so that we can obtain insight into what the functional landscape looks like. (iii) Establishing a protein fate map that shows statistically what happens to proteins in cells so that we can determine what the impacts of cellular changes are. (iv) Determining the numbers that govern protein production and verifying them by calculation. Indicating that there are still major gaps in our knowledge of the protein production numbers and that the era of quantitative biology is still in its early phase.

Acknowledgments

The work was funded by an NWO TOP-GO program grant (#L.10.060) and an ERC Advanced Grant (ABCVolume, #670578).

References

- Moore, P.B. (2012) How should we think about the ribosome? *Annu. Rev. Biophys.* 41, 1–19
- Nijeholt, J.A.L.A. and Driessen, A.J.M. (2012) The bacterial Sec translocase: structure and mechanism. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 367, 1016–1028
- Jomaa, A. *et al.* (2017) Structure of the quaternary complex between SRP, SR, and translocon bound to the translating ribosome. *Nat. Commun.* 8, 15470
- Nowak, M. (2006) Evolutionary dynamics. In *Evolutionary Dynamics*. Belknap Press, pp. 12
- Milo, R. and Phillips, R. (2016) *Cell Biology by the Numbers*. Garland Science, pp. 11
- Young, K.D. (2006) The selective value of bacterial shape. *Microbiol. Mol. Biol. Rev.* 70, 660–703
- Osterman, I.A. *et al.* (2015) Regulation of flagellar gene expression in bacteria. *Biochemistry (Moscow)* 80, 1447–1456
- Dix, J.A. and Verkman, A.S. (2008) Crowding effects on diffusion in solutions and cells. *Annu. Rev. Biophys.* 37, 247–263
- Zhou, H. *et al.* (2008) Macromolecular crowding and confinement: biochemical, biophysical, and potential physiological consequences. *Annu. Rev. Biophys.* 37, 375–397
- van den Berg, J. *et al.* (2017) Microorganisms maintain crowding homeostasis. *Nat. Rev. Microbiol.* 15, 309–318
- Taniguchi, Y. *et al.* (2010) Quantifying *E. coli* proteome and transcriptome with single-molecule sensitivity in single cells. *Science* 329, 533–538
- Bakshi, S. *et al.* (2012) Superresolution imaging of ribosomes and RNA polymerase in live *Escherichia coli* cells. *Mol. Microbiol.* 85, 21–38
- Moffitt, J.R. *et al.* (2016) Spatial organization shapes the turnover of a bacterial transcriptome. *Elife* 5, e13065
- Toro, E. and Shapiro, L. (2010) Bacterial chromosome organization and segregation. *Cold Spring Harb. Perspect. Biol.* 2, a000349
- Shieh, Y. *et al.* (2015) Operon structure and cotranslational subunit association direct protein assembly in bacteria. *Science* 350, 678–680
- Phillips, R. *et al.* (2009) *Physical Biology of the Cell*, pp. 481–512. Garland Science
- Andrews, S.S. *et al.* (2010) Detailed simulations of cell biology with Smoldyn 2.1. *PLoS Comput. Biol.* 6, e1000705
- Chong, S. *et al.* (2014) Mechanism of transcriptional bursting in bacteria. *Cell* 158, 314–326
- Li, G. and Xie, X.S. (2011) Central dogma at the single-molecule level in living cells. *Nature* 475, 308–315
- Chen, H. *et al.* (2015) Genome-wide study of mRNA degradation and transcript elongation in *Escherichia coli*. *Mol. Syst. Biol.* 11, 781
- Yan, X. *et al.* (2016) Dynamics of translation of single mRNA molecules *in vivo*. *Cell* 165, 976–989
- Geisberg, J.V. *et al.* (2014) Global analysis of mRNA isoform half-lives reveals stabilizing and destabilizing elements in yeast. *Cell* 156, 812–824
- Wen, J. *et al.* (2008) Following translation by single ribosomes one codon at a time. *Nature* 452, 598–603
- Gardin, J. *et al.* (2014) Measurement of average decoding rates of the 61 sense codons *in vivo*. *Elife* 3, e03735
- Janssen, B.D. and Hayes, C.S. (2012) The tmRNA ribosome-rescue system. *Adv. Protein Chem. Struct. Biol.* 86, 151–191
- Bai, L. *et al.* (2006) Single-molecule analysis of RNA polymerase transcription. *Annu. Rev. Biophys. Biomol. Struct.* 35, 343–360
- van Oijen, A.M. and Loparo, J.J. (2010) Single-molecule studies of the replisome. *Annu. Rev. Biophys.* 39, 429–448
- Kumazaki, K. *et al.* (2014) Structural basis of Sec-independent membrane protein insertion by YidC. *Nature* 509, 516–520
- Wiedemann, N. and Pfanner, N. (2017) Mitochondrial machineries for protein import and assembly. *Annu. Rev. Biochem.* 86, 685–714

Outstanding Questions

How, and to what extent, do the various levels of the protein production process connect? Which features of higher levels are (in)dependent on the nature of the lower levels?

What does the functional landscape of protein production look like and how does it connect to other cellular processes? Why does the Sec translocon perform both insertion into the membrane and secretion across it?

What is the life history of individual proteins and how does this depend on the action of chaperones and proteases? Is it possible to have the same functional capacities in cells without chaperones and proteases?

What are the numbers, concentrations, rates, diffusion coefficients, etc. associated with (membrane) protein production? Which of these numbers can be connected by calculation so that they can be verified and used to make predictions?

30. Goossens, V.J. *et al.* (2014) The Tat system of Gram-positive bacteria. *Biochim. Biophys. Acta* 1843, 1698–1706
31. Patel, S. and Picha, K. (2000) Structure and function of hexameric helicases. *Annu. Rev. Biochem.* 69, 651–697
32. van Gijtenbeek, L.A. *et al.* (2016) On the spatial organization of mRNA, plasmids, and ribosomes in a bacterial host overexpressing membrane proteins. *PLoS Genet.* 12, e1006523
33. Franken, L.E. *et al.* (2017) A general mechanism of ribosome dimerization revealed by single-particle cryo-electron microscopy. *Nat. Commun.* 8, 722
34. Schavemaker, P.E. *et al.* (2017) Ribosome surface properties may impose limits on the nature of the cytoplasmic proteome. *Elife* 6, e30084
35. Milo, R. *et al.* (2010) BioNumbers – the database of key numbers in molecular and cell biology. *Nucleic Acids Res.* 38, D750–D753
36. Milo, R. and Phillips, R. (2016) *Cell Biology by the Numbers*, Garland Science
37. Keseler, I.M. *et al.* (2017) The EcoCyc database: reflecting new knowledge about *Escherichia coli* K-12. *Nucleic Acids Res.* 45, D543–D550
38. Krogh, A. *et al.* (2001) Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J. Mol. Biol.* 305, 567–580
39. Vendeville, A. *et al.* (2011) An inventory of the bacterial macromolecular components and their spatial organization. *FEMS Microbiol. Rev.* 35, 395–414
40. Milo, R. (2013) What is the total number of protein molecules per cell volume? A call to rethink some published values. *Bioessays* 35, 1050–1055
41. Goodsell, D.S. (2009) *The Machinery of Life*, Springer, pp. 6
42. Milo, R. and Phillips, R. (2016) *Cell Biology by the Numbers*, Garland Science, pp. 232
43. Maurizi, M. (1992) Proteases and protein-degradation in *Escherichia coli*. *Experientia* 48, 178–201
44. Wang, M. *et al.* (2012) PaxDb, a database of protein abundance averages across all three domains of life. *Mol. Cell. Proteomics* 11, 492–500
45. Marintchev, A. and Wagner, G. (2004) Translation initiation: structures, mechanisms and evolution. *Q. Rev. Biophys.* 37, 197–284
46. Voorhees, R.M. and Ramakrishnan, V. (2013) Structural basis of the translational elongation cycle. *Annu. Rev. Biochem.* 82, 203–236
47. Woese, C. *et al.* (2000) Aminoacyl-tRNA synthetases, the genetic code, and the evolutionary process. *Microbiol. Mol. Biol. Rev.* 64, 202–236
48. Driessen, A.J.M. and Nouwen, N. (2008) Protein translocation across the bacterial cytoplasmic membrane. *Annu. Rev. Biochem.* 77, 643–667
49. Kim, Y.E. *et al.* (2013) Molecular chaperone functions in protein folding and proteostasis. *Annu. Rev. Biochem.* 82, 323–355
50. Gur, E. *et al.* (2011) Regulated proteolysis in Gram-negative bacteria – how and when? *Nat. Rev. Microbiol.* 9, 839–848
51. Yu, Z. *et al.* (2011) Role for *Escherichia coli* YidD in membrane protein insertion. *J. Bacteriol.* 193, 5242–5251
52. Ito, K. and Akiyama, Y. (2005) Cellular functions, mechanism of action, and regulation of FtsH protease. *Annu. Rev. Microbiol.* 59, 211–231
53. Milo, R. and Phillips, R. (2016) *Cell Biology by the Numbers*, Garland Science, pp. 10
54. Linden, M. *et al.* (2012) Entropic tension in crowded membranes. *PLoS Comput. Biol.* 8, e1002431
55. von Heijne, G. (2006) Membrane-protein topology. *Nat. Rev. Mol. Cell Biol.* 7, 909–918