Mass distributions of a macromolecular assembly based on electrospray ionization mass spectrometric masses of the constituent subunits

LEONID HANIN\textsuperscript{a}, BRIAN GREEN\textsuperscript{b}, FRANCK ZAL\textsuperscript{c} and SERGE VINOGRADOV\textsuperscript{d, *}

\textsuperscript{a}Department of Mathematics, Idaho State University, Pocatello, ID 83209, USA
\textsuperscript{b}Micromass UK Ltd., Altrincham, Cheshire, WA14 5RZ, UK
\textsuperscript{c}Equipe d’Ecophysiologie, Station Biologique, UPMC-CNRS-INSU, 29682 Roscoff, France
\textsuperscript{d}Department of Biochemistry and Molecular Biology, Wayne State University School of Medicine, Detroit, MI 48201, USA

*Corresponding author (Fax, 313-577-2765; Email, svinogra@med.wayne.edu)

Macromolecular assemblies containing multiple protein subunits and having masses in the megadalton (MDa) range are involved in most of the functions of a living cell. Because of variation in the number and masses of subunits, macromolecular assemblies do not have a unique mass, but rather a mass distribution. The giant extracellular erythrocruorins (Ers) of annelids are giant, ~ 3500 kDa, complexes which were among the first protein assemblies investigated by ultracentrifugation (Svedberg 1933) and electron microscopy (EM) (Levin 1963; Roche 1965). Their characteristic EM and scanning transmission electron microscopy (STEM) appearances are those of an hexagonal bilayer (HBL) with diameter ~ 30 nm and height ~ 20 nm (Roche 1965; Terwilliger et al. 1976; Vinogradov et al. 1982; Boekema and van Heel 1989; Gotoh and Suzuki 1990; Lamy et al. 1996). Extensive small angle X-ray scattering (SAXS)

1. Introduction

Most of the functions in a living cell are carried out by “molecular machines”, large macromolecular assemblies comprised of numerous and different proteins, with masses in the MDa range (Alberts 1998; Nogales and Grigorieff 2001). The variation in the number and masses of the constituent subunits imply that large protein assemblies do not have a unique mass, but rather a mass distribution. The extracellular erythrocruorins (Ers) of annelids are giant, ~ 3500 kDa, complexes which were among the first protein assemblies investigated by ultracentrifugation (Svedberg 1933) and electron microscopy (EM) (Levin 1963; Roche 1965). Their characteristic EM and scanning transmission electron microscopy (STEM) appearances are those of an hexagonal bilayer (HBL) with diameter ~ 30 nm and height ~ 20 nm (Roche 1965; Terwilliger et al. 1976; Vinogradov et al. 1982; Boekema and van Heel 1989; Gotoh and Suzuki 1990; Lamy et al. 1996). Extensive small angle X-ray scattering (SAXS)

Keywords. Electrospray ionization mass spectrometry; erythrocruorin; lumbricus; mass distribution; riftia

Abbreviations used: cryoEM, cryoelectron microscopy; 3D, three-dimensional; E, mathematical expectation; EM, electron microscopy; Er, erythrocruorin; ESI-MS, electrospray ionization mass spectrometric; HBL, hexagonal bilayer; pdf, probability density function; rv, random variable; SAXS, small angle X-ray scattering; SD, standard deviation; STEM, scanning transmission electron microscopy.
studies have shown the molecular shapes and dimensions of different Ers to be very similar (Stöckel et al 1973; Pilz et al 1980, 1986, 1988; Wilhelm et al 1980; Messerschmidt et al 1983; Terwilliger and Terwilliger 1985; Theuer et al 1985; Krebs et al 1998). More recently, three-dimensional (3D) reconstructions using cryoelectron microscopy (cryoEM) have demonstrated that all the quaternary structures are virtually identical at a resolution of ~2 nm (Cejka et al 1989, 1991, 1992; De Haas et al 1996a, b, c, d, 1997; Taveau et al 1999; Jouan et al 2001).

Concurrent electrospray ionization mass spectrometric (ESI-MS) studies provided accurate masses for the constituent globin and linker chains and the disulphide-bonded globin dimer/trimer/tetramer subunits (Green et al 1995; 1998a, 1999; Weber et al 1995; Martin et al 1996; Zal et al 1996, 1997a, b, 2000). The very recent low resolution crystallographic structure of Lumbricus Er (Royer et al 2000) established the stoichiometry to be 144 globin chains arranged as 12 dodecamers tethered to 36 linker chains, in agreement with the model proposed earlier (Vinogradov et al 1986) and the cryoEM 3D reconstructions (De Haas et al 1996a, b, c, d, 1997; Taveau et al 1999; Jouan et al 2001). In sharp contrast to the evidence for very similar molecular dimensions and quaternary structures, the over 60 known Er masses, obtained by several experimental techniques from the early work of Svedberg to the present, exhibit a surprisingly wide range, i.e. ranging from 2400 to 4470 kDa (Lamy et al 1996).

Because the 144 globin chains consist of 4 to 11 different proteins with masses from 16 to 19 kDa, and the 36 linker chains also consist of 2 to 7 different proteins with masses from 24 to 32 kDa (Lamy et al 1996), Er, like all large molecular assemblies, are expected to have a mass distribution rather than a unique mass. There are four levels of mass distribution possible: the dodecamer subassembly of globin chains; the complex of 12 dodecamer subassemblies; the linker subassembly of 36 linker chains which are required for HBL structure formation (Kuchumov et al 1999); and finally the overall mass distribution of the complete assembly. We report below the results of calculations of mass distributions for the subassemblies and complete assemblies of the two Ers which differ in the subunit nature of their dodecamer subassemblies: from the earthworm Lumbricus; and, from the deep-sea polychaete Riftia.

2. Methods

2.1 Mass distributions for a single dodecamer subassembly

2.1a Lumbricus: Each dodecamer subassembly consists of 3 monomers which are randomly chosen from 3 different globin chains d1, d2, d3 and 3 trimers randomly selected from the observed 4 trimers t1-t4 (Martin et al 1996). We assume that the monomeric and trimeric subunits are incorporated in a dodecamer independently of each other. Let \( p(dt), t = 1, 2, 3, \) and \( p(tj), j = 1, 2, 3, 4, \) be the probabilities that the corresponding monomer chains and trimers are present in a dodecamer. Let \( D = (D_1, D_2, D_3) \) be the vector of random numbers of monomers \( d_1, d_2, d_3 \) present in a dodecamer, subject to the condition \( D_1 + D_2 + D_3 = 3 \). The random vector \( D \) takes 10 values that correspond to the following combinations of monomer subunits: \( 3d_1, 3d_2, 3d_3, 2d_1d_2, 2d_2d_3, 2d_1d_3, 2d_3d_1, 2d_2d_1, d_1d_2d_3 \). Consequently, the random vector \( D \) has the multinomial distribution \( B[3; p(d_1), p(d_2), p(d_3)] \), i.e.

\[
Pr(\delta) := Pr(D = \delta) = \frac{3!}{\delta_1!\delta_2!\delta_3!} p(d_1)^{\delta_1} p(d_2)^{\delta_2} p(d_3)^{\delta_3},
\]

where \( \delta = (\delta_1, \delta_2, \delta_3) \) is a vector of nonnegative integers such that \( \delta_1 + \delta_2 + \delta_3 = 3 \).

Similarly, let \( T = (T_1, T_2, T_3, T_4) \) be the vector of random numbers of trimers t1-t4 contained in a dodecamer. Obviously, \( T_j, 1 \leq j \leq 4, \) are nonnegative integral-valued random variables such that \( \sum_{j=1}^4 T_j = 3 \). The random vector \( T \) takes 20 values that correspond to the following combinations of trimer subunits: \( 3t_1, 3t_2, 3t_3, 3t_4, 2t_1t_2, 2t_2t_1, 2t_1t_3, 2t_3t_1, 2t_1t_4, 2t_4t_1, 2t_2t_3, 2t_3t_2, 2t_2t_4, 2t_4t_2, 2t_3t_4, 2t_4t_3, t_1t_2t_3, t_1t_2t_4, t_1t_3t_4, t_2t_3t_4 \). Consequently, the random vector \( T \) follows the multinomial distribution \( B[3; p(t_1), p(t_2), p(t_3), p(t_4)] \), i.e.

\[
p(\tau) := Pr(T = \tau) = 3! \prod_{j=1}^4 \frac{p(t_j)^{\tau_j}}{\tau_j!},
\]

where \( \tau = \{ \tau : 1 \leq j \leq 4 \} \) is a vector of nonnegative integers such that \( \sum_{j=1}^4 \tau_j = 3 \). Thus, the number of different Lumbricus dodecamer subassemblies is equal to \( 10 \times 20 = 200 \). Each of them is determined by the two integer vectors \( \delta \) and \( \tau \) described above, and its probability is equal to the product of probabilities (1) and (2). The mass of a randomly assembled dodecamer is given by:

\[
m = \sum_{i=1}^3 m(d_i)D_i + \sum_{j=1}^4 m(t_j)T_j + h,
\]

where \( h = 12 \times 616.5 = 7398 \) Da is the total heme mass, and the observed electrospray ionization mass spectrometric masses of monomers and trimers are \( m(d_1) = 15993 \) Da, \( m(d_2) = 15978 \) Da, \( m(d_3) = 15962 \) Da and \( m(t_1) = 52923 \) Da, \( m(t_2) = 52760 \) Da, \( m(t_3) = 52599 \) Da, \( m(t_4) = 52435 \) Da, with probabilities derived from the observe intensities of the peaks, \( p(d_1) = 0.55, p(d_2) = 0.28, p(d_3) \)