Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

Comment on *Microfibrillar structure of type I collagen in situ* by Orgel *et al.* (2006), *Proc. Natl Acad. Sci. USA*, 103, 9001–9005

Kenji Okuyama,[a]* Hans Peter Bächinger,[b] Kazunori Mizuno,[b] Sergei Boudko,[b] Jürgen Engel,[c] Rita Berisio[d] and Luigi Vitagliano[d]

[a]Department of Macromolecular Science, Graduate School of Science, Osaka University, Toyonaka, Osaka 560-0043, Japan, [b]Research Department, Shriners Hospital for Children, Portland, OR 97239, USA, [c]Biozentrum, University of Basel, Klingelbergstrasse 70, CH-4056 Basel, Switzerland, and [d]Istituto Biostruttre e Bioimmagini, CNR, via Mezzocannone 16, I-80134 Napoli, Italy

Correspondence e-mail: okuyamak@chem.sci.osaka-u.ac.jp

Received 23 March 2009
Accepted 9 June 2009

A comment is published on the article *Microfibrillar structure of type I collagen in situ* by Orgel *et al.* [(2006), *Proc. Natl Acad. Sci. USA*, **103**, 9001–9005].

The molecular structure of collagen represents a long-standing issue in structural biology. The complexity and the fibrous nature of the protein prevent the application of single-crystal crystallographic techniques. Although partial information on the structure of collagen has been derived by using polypeptide models, the structural characterization of the full-length protein would provide an invaluable tool for understanding the many biological processes in which collagen is involved. The determination of the molecular structure of collagen from wide-angle X-ray fiber diffraction data has also proven to be extremely difficult, despite the progress of fiber diffraction techniques over the last eight decades. Because of a deficiency of diffraction spots on the layer lines in the wide-angle region (*ca* 1–30 Å resolution), it could not even be determined whether the average helical symmetry of the collagen superhelix was 7/2 (seven tripeptide units per two turns) or 10/3 (Okuyama *et al.*, 2006). In a recently published article, *Microfibrillar structure of type I collagen in situ* (Orgel *et al.*, 2006), the authors report the three-dimensional molecular and packing structure of type I collagen determined by X-ray fiber diffraction analysis, which was based on 414 reflections with a completeness of 5% in the range of 5–113 Å resolution (PDB entry 1y0f). The collagen molecule is made of three chains of more than 1000 residues each. Can we determine the three-dimensional molecular conformation based on such a small number of reflections at low resolution? Most readers would be likely to fall under this impression. However, because the fiber diffraction analysis combined with heavy-atom isomorphous replacement is a highly specialized methodology, almost all readers of Orgel's paper (including the authors of this letter initially) took their results at face value. Orgel's structure has been referenced by many researchers as the molecular structure of the collagen fibril. Furthermore, this paper was nominated as a paper of outstanding interest in recent reviews (Tsuruta & Irving, 2008; Vakonakis & Campbell, 2007).

Recently, we carefully analyzed the PDB entry 1y0f to evaluate the helical symmetry of collagen α-chains in Orgel's model. Although, as observed for most collagen-like peptides, the average helical symmetry of Orgel's model is 7/2-helix, we found some questionable aspects in their analysis.

(1) *Chain sequence*. Orgel *et al.* collected fiber diffraction data from rat-tail tendon collagen, and cited SwissProt acquisition codes P02454 and P02466 in the deposited data (1y0f) as the amino-acid sequences of α1(I) and α2(I) chains, respectively. It followed from a biochemical analysis, that collagen was present in its enzymatically processed tissue form. Strangely, the sequence used for the structure derived by Orgel *et al.* differs substantially from the cited code. For the α1(I) chain, their deposited sequence has 39 differences relative to P02454, including two missing residues at the C-terminus. In the α2(I) chain, there are 147 differences, including two missing residues in the N-terminal telopeptide, three missing residues between 876 and 877, Gly-Ala-Ala in P02466, and the last nine missing residues at the end

of the C-terminus. (The numbers were calculated with the assumption that processing of type I procollagen in rat tail tendon is similar to that in the other tissues.)

(2) *Chain arrangement*. In the collagen helix, each peptide chain must be staggered by one residue with respect to its neighbor, in order to ensure that every glycine in the sequence is available to localize near the common axis. Since type I collagen is a heterotrimer composed of two α1(I) chains and one α2(I) chain, there are three possible arrangements, α1(I)α1(I)α2(I), α1(I)α2(I)α1(I) and α2(I)α1(I)α1(I). We understand that the actual arrangement not yet been solved, however, Orgel *et al.* used the second arrangement in most of the molecule without offering any justifying explanation. Their assumption could have been proven by refining three distinct models with the α2(I) chain located in different positions. This check would have also provided insights into the possibility of discriminating correct *versus* incorrect models with the available experimental data. Furthermore, a tripeptide is missing between residues 876 and 877 of the α2(I) chain. This leap in the sequence should have a twofold consequence: (i) it should cause a different chain order from this location to the C terminus and (ii) it should cause a drastic change in the telopeptide conformation.

(3) *Residue occupancy*. Although Orgel *et al.* used fixed temperature factors for Cα atoms, the occupancies of 2517 residues (out of 3134) are not 1.0. For example, out of 2517, 134 residues have occupancy factors as small as 0.15, which means only 15% of these sites are occupied. Of course, the temperature factor and occupancy of a given atom are mutually related. However, it is not reasonable to change residue occupancies in order to obtain good agreement between observed and calculated structure amplitudes because of the limited number of available experimental data at low resolution.

(4) *Data/parameter ratio*. In fiber diffraction analyses of crystalline polymers (including DNA, polysaccharides, and synthetic polymers), the linked-atom least-squares (LALS) method (Arnott & Wonacott, 1966; Smith & Arnott, 1978) has been the most well known for solving molecular and packing structures based on the fiber diffraction data in the wide-angle region. The molecular structure of collagen was analyzed using this method (Fraser *et al.*, 1979; Okuyama *et al.*, 2006). It was also used for the single-crystal analysis of a collagen-model peptide, using 401 unique reflections with a completeness of 51% up to 2.2 Å resolution (Okuyama *et al.*, 1981). In the LALS method, the refinement parameters are basically conformation angles in a helical repeating unit, together with positioning and orienting parameters that locate and orient the polymer chain in its unit cell. The values of bond lengths and bond angles are usually fixed to their standard values, in order to decrease the number of refinement parameters; this compensates for the deficiency of diffraction data in the fiber diffraction patterns. Furthermore, instead of refining temperature factors of all atoms, only one overall temperature factor is refined. In this way, the ratio of observed data (401) and variable parameters (26) became reasonable (Okuyama *et al.*, 1981). In the analysis of Orgel *et al.,* judging from deposited values and *Supporting Methods*, occupancy factors were refined for 3000 residues, and backbone and side-chain atoms were included in the refinement (http://www.pnas.org/content/103/24/9001/suppl/DC2). This procedure is rather singular, if it is considered that parameters were refined against the observed 414 reflections. Consequently, the credibility of the obtained model should be considered to be very low.

(5) *The collagen structure: a three-dimensional model to be handled with care*. The dissemination of protein three-dimensional models through structural databases such as the Protein Data Bank (Berman *et al.*, 2002) has broadened the impact of structural biology studies, by stimulating an enormous number of structure-based biochemical and biological experiments. The availability of protein three-dimensional models to biologically oriented communities, however, presents some drawbacks. Indeed, it is not obvious to all users that the deposited protein structures are, in principle, only models used to interpret the actual experimental data, *i.e.* the diffraction pattern. Even the overall correctness of the structure does not guarantee the accuracy of specific protein regions.

In conclusion, the points raised here indicate that the structure of collagen presented by Orgel and coworkers should be handled with care. Indeed, although the triple helix tracing may be correct, the assignment of the sequence to their model and, therefore, the positioning of the two α1(I) and α2(I) chains remain ambiguous. We hope that the present comment will stimulate a debate on a crucial issue of the current understanding of the collagen structure.

References

Arnott, S. & Wonacott, A. J. (1966). *J. Mol. Biol.* **21**, 371–383.

Berman, H. M., Battistuz, T., Bhat, T. N., Bluhm, W. F., Bourne, P. E., Burkhardt, K., Feng, Z., Gilliland, G. L., Iype, L., Jain, S., Fagan, P., Marvin, J., Padilla, D., Ravichandran, V., Schneider, B., Thanki, N., Weissig, H., Westbrook, J. D. & Zardecki, C. (2002). *Acta Cryst.* D**58**, 899–907.

Fraser, R. D., MacRae, T. P. & Suzuki, E. (1979). *J. Mol. Biol.* **129**, 463–481.

Okuyama, K., Okuyama, K., Arnott, S., Takayanagi, M. & Kakudo, M. (1981). *J. Mol. Biol.* **152**, 427–443.

Okuyama, K., Xu, X., Iguchi, M. & Noguchi, K. (2006). *Biopolymers*, **84**, 181–191.

Orgel, J. P. R. O., Irving, T. C., Miller, A. & Wess, T. J. (2006). *Proc. Natl Acad. Sci. USA*, **103**, 9001–9005.

Smith, P. J. C. & Arnott, S. (1978). *Acta Cryst.* A**34**, 3–11.

Tsuruta, H. & Irving, T. C. (2008). *Curr. Opin. Struct. Biol.* **18**, 601–608.

Vakonakis, I. & Campbell, I. D. (2007). *Curr. Opin. Cell Biol.* **19**, 578–583.

Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

On the packing structure of collagen: response to Okuyama *et al.*'s comment on *Microfibrillar structure of type I collagen in situ*

Joseph. P. R. O. Orgel

BioCAT and μCoSM Centres: Pritzker Institute of Biomedical Science and Engineering, Illinois Institute of Technology, 3440 S. Dearborn Avenue, Chicago, IL 60616, USA, and CSRRI and Department of Biological, Chemical and Physical Sciences, Illinois Institute of Technology, 3101 S. Dearborn Ave, Chicago, IL 60616, USA

Correspondence e-mail: orgel@iit.edu

Received 15 June 2009
Accepted 15 June 2009

A response is published to the comment by Okuyama *et al.* [(2009) *Acta Cryst.* D**65**, 1007–1008] on *Microfibrillar structure of type I collagen in situ*.

1. Introduction

It is disappointing to us that Okuyama *et al.* (2009) chose to largely ignore the most important and substantially supported aspects of our study, namely collagen's molecular packing structure. Instead, by either misunderstanding or through selective attention, they present minor flaws in the coordinate file 1y0f as if they are serious blows to the overall study.

2. The first experimentally determined (low-resolution) packing structure of collagen

The purpose of Orgel *et al.* (2006) was to determine the relative spatial arrangement of the five collagen molecules in the unit cell of natively crystalline rat-tail tendon without a dependency on experimentally biased models. This was an essential first step before more detailed structural models could build upon, improve or surpass the initial work. The electron-density map, constructed from experimentally determined phases and observed amplitudes, is clearly and prominently shown and compared with the low-resolution and coordinate based models [see *Supporting Methods* published as supporting information (SI) in Orgel *et al.* (2006)] and $2F_o - F_c$ electron-density map in the paper, and all show good agreement. Hence, at the resolution of the study (5.16 Å axial and 11.1 Å equatorial) we stand by its conclusions.

As a byproduct of the final steps in our attempt to exhaustively test the accuracy of the experimental results (SI Table 3, *Supporting Methods* of Orgel, 2006), the coordinates contained in 1y0f and 1ygv were reached by fitting high-resolution collagen-like peptide structural data into our low-resolution electron-density map, essentially a molecular envelope. This approach is analogous with 'docking' fragments of a high-resolution structure into low-resolution molecular envelopes derived from cryo-electron microscopy or SAXS data (Henderson, 2004; Petoukhov & Svergun, 2007). These represent credible attempts to establish the context in which these detailed, but incomplete, pieces of the puzzle fit together. No-one should confuse the resulting small-scale features of those fragments within the low-resolution structures with those derived by high-resolution single-crystal crystallography or multidimensional NMR. In our case, the low-resolution molecular envelope details the gross arrangement of the collagen molecules, and is not suitable for the study of the specific helical conformation, without further higher resolution equatorial data.

In communicating the coordinate files to the RCSB database, it was our hope that these would provide useful starting points for subsequent studies. At the same time, our caution and transparency in submitting both the 'rigid' (1ygv) and 'relaxed' (1y0f) models and only the Cα atoms in both should communicate clearly that the

coordinates are derived from low-resolution data and should be handled appropriately. This point is further made by the fiber diffraction specific annotations within the files and the substantial SI material contributed with the original publication showing what was done and how.

3. Specific issues

3.1. Completeness

Okuyama *et al.* (2009) misinterpret the information within the 1y0f and 1ygv coordinate files. By mistaking the resolution of the study as isotropic, they assume that 5% represents the completeness of the whole data set. This is despite the fact that in both Orgel *et al.* (2006) and the RCSB coordinate files the resolution is clearly shown to be of anisotropic resolution (5.16 Å axial and 11.1 Å equatorial). Both the publication and coordinate files discuss the number of observed and utilized reflections and the completeness of the refinement data set is actually around 95%.

3.2. Chain sequence

The chain sequences were mostly right. The discrepancies between the coordinate file sequence [linked to earlier studies (Orgel *et al.*, 2000) when the sequence at the end of the α2 sequence was uncertain] and the updated Uniprot data are a small percentage of the whole molecule and do not effect chain registration *etc.* The comment that nine residues are missing from the C-terminus of the α2 sequence seems to be incorrect as we understand the rat α2 C-terminal region to be shorter than that of other species and the other telopeptide differences were trivial, but we thank Okuyama *et al.* for bringing these to our attention.

More importantly however, it should be noted that given the resolution of the study and given that only Cα positions were reported, these errors are of little or no significance; any mammalian type I collagen sequence would have sufficed for the purpose of model refinement. In our case, after repeating the refinement of the molecular packing model with the corrected sequences, we found no change in the molecular trace, only trivial changes in the specific peptide chain position and no significant change in the *R* factors (or *b*/*q* factors). The small reduction in *R* factor with the corrected sequence indicates that the refinement method is fundamentally sound. We have uploaded the sequence corrected files as referenced under RCSB codes 3hqv and 3hr2.

3.3. Chain arrangement

The peptide chain registration, the position of the whole helices relative to the electron density, cross-linking locations and telopeptide conformations were based on the alignment shown in Orgel *et al.* (2000) and Orgel *et al.* (2001), which were referenced in Orgel *et al.* (2006). Here, the heavy atoms in isomorphous derivatives serve as markers of key sequence elements (*e.g.* the Tyr residues in the telopeptides). These features are in no way dependent on the 1ygv or 1y0f models; they were determined independently of them. Rather, the models were constructed to include these experimentally observed features.

3.4. Residue occupancy *versus* temperature factor

Okuyama *et al.* raise an important concern, but the regional calculation of temperature factor and lattice distortions were, in fact, discussed in Orgel *et al.* (2006): the temperature factor was assessed as 190 Å^2 for the molecule overall. The use of the '*q* factor' was clearly stated in the publication and what its relation is to the overall temperature factor. It does not refer to the residue 'absence' in our study. In the refinement of the coordinate models, we chose to use the *q* factor as a more parsimonious approach because both *q* and *b* factors are approximations and either parameter has roughly equivalent effects at this resolution and we did not refine >3000 parameters at the same time (see SI *Supporting Methods*). What is more, the low-resolution pre-refinement model used only a handful of regional (along the D-period/crystallite *c* axis) temperature factors and the fit of the sequence to the data was good (initial model in SI *Supporting Methods* and SI Fig. 12).

3.5. Data-to-parameter ratio

In the *Supporting Methods* to Orgel *et al.* (2006) it is clearly explained that there was an approximately tenfold excess of data to parameters in the refinement of the 1ygv coordinates and how this was achieved. For instance, rather than refining the individual position of 3300 amino-acid residues, the molecular refinement involved

> ... defining 46 regions of the collagen triple helix that are relatively straight, as individual rigid bodies of different lengths, connected by short sections (average length »6 aa) of triple helix that were not constrained, the latter acting as hinges for the refinement of the straight sections. This greatly restrained the degrees of freedom involved in the molecular refinement ...

The final coordinates in 1y0f did not have this degree of constraint, but the molecular trace does not deviate significantly from that of 1ygv. The significance of this last step was that only the stereochemistry of the bonds and the experimental electron density constrained the fit, allowing for some insight into how disassociated the peptide chains might be from the triple-helix in some parts of the molecular packing structure. This is seen in the varying diameter of the electron-density 'tubes' showing the outline of the collagen molecules.

3.6. The collagen structure, a model to be handled with care

The coordinates we have contributed currently represent the best known alignment of collagen sequence to the three-dimensional packing structure of collagen molecules *in situ*, despite their known deficiencies. They are not, and were never intended to be a direct contribution to our understanding of collagen's triple-helical symmetry as Okuyama *et al.* appear to believe. However, we fully agree with Okuyama *et al.*'s conclusion that the coordinates provided in Orgel *et al.* (2006) should be used with care and with due consideration of their intrinsic limitations.

References

Henderson, R. (2004). *Q. Rev. Biophys.* **37**, 3–13.

Okuyama, K., Bachingera, H. P., Mizuno, K., Boudko, S., Engel, J., Berisio, R. & Vitagliano, L. (2009). *Acta Cryst.* D**65**, 1007–1008.

Orgel, J. P., Irving, T. C., Miller, A. & Wess, T. J. (2006). *Proc. Natl Acad. Sci. USA*, **103**, 9001–9005.

Orgel, J. P., Miller, A., Irving, T. C., Fischetti, R. F., Hammersley, A. P. & Wess, T. J. (2001). *Structure*, **9**, 1061–1069.

Orgel, J., Wess, T. & Miller, A. (2000). *Structure Fold. Des.* **8**, 137–142.

Petoukhov, M. & Svergun, D. (2007). *Curr Op. Struct. Biol.* **17**, 562–571.