

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**

^aDepartment of Macromolecular Science, Graduate School of Science, Osaka University, Toyonaka, Osaka 560-0043, Japan, ^bResearch Department, Shriners Hospital for Children, Portland, OR 97239, USA, ^cBiozentrum, University of Basel, Klingelbergstrasse 70, CH-4056 Basel, Switzerland, and ^dIstituto Biostrutture e Bioimmagini, CNR, via Mezzocannone 16, I-80134 Napoli, Italy

Correspondence e-mail:

okuyamak@chem.sci.osaka-u.ac.jp

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].

Received 23 March 2009

Accepted 9 June 2009

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 $\alpha 1(I)$ and $\alpha 2(I)$ 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 $\alpha 1(I)$ chain, their deposited sequence has 39 differences relative to P02454, including two missing residues at the C-terminus. In the $\alpha 2(I)$ 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 $\alpha 1(\text{I})$ chains and one $\alpha 2(\text{I})$ chain, there are three possible arrangements, $\alpha 1(\text{I})\alpha 1(\text{I})\alpha 2(\text{I})$, $\alpha 1(\text{I})\alpha 2(\text{I})\alpha 1(\text{I})$ and $\alpha 2(\text{I})\alpha 1(\text{I})\alpha 1(\text{I})$. 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 $\alpha 2(\text{I})$ 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 $\alpha 2(\text{I})$ 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 C α 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 $\alpha 1(\text{I})$ and $\alpha 2(\text{I})$ 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.* **D58**, 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.* **A34**, 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.