Atti Accademia Nazionale dei Lincei

Classe Scienze Fisiche Matematiche Naturali

Rendiconti

Alfonso M. Liquori, Alberto Ripamonti, Claudia Sadun, Stefano Ottani, Dario Braga

Fourier analysis of the primary structure of globular proteins

Atti della Accademia Nazionale dei Lincei. Classe di Scienze Fisiche, Matematiche e Naturali. Rendiconti, Serie 8, Vol. **75** (1983), n.1-2, p. 71–78.

Accademia Nazionale dei Lincei

<http://www.bdim.eu/item?id=RLINA_1983_8_75_1-2_71_0>

L'utilizzo e la stampa di questo documento digitale è consentito liberamente per motivi di ricerca e studio. Non è consentito l'utilizzo dello stesso per motivi commerciali. Tutte le copie di questo documento devono riportare questo avvertimento.

Articolo digitalizzato nel quadro del programma bdim (Biblioteca Digitale Italiana di Matematica) SIMAI & UMI
http://www.bdim.eu/

Atti della Accademia Nazionale dei Lincei. Classe di Scienze Fisiche, Matematiche e Naturali. Rendiconti, Accademia Nazionale dei Lincei, 1983.

SEZIONE II

(Fisica, chimica, geologia, paleontologia e mineralogia)

Chimica. — *Fourier analysis of the primary structure of globular proteins.* Nota di ALFONSO M. LIQUORI (*), ALBERTO RIPAMONTI (**), CLAUDIA SADUN (***), STEFANO OTTANI (****) e DARIO BRAGA (**), presentata (*****) dal Socio V. CAGLIOTI.

RIASSUNTO. — Viene introdotto un nuovo algoritmo per analizzare la struttura primaria (cioè la sequenza degli amminoacidi) di una proteina globulare. Esso si basa sul calcolo di una funzione di autocorrelazione vettoriale come serie di Fourier analoga alla funzione Patterson impiegata nell'analisi di strutture cristalline mediante diffrazione dei raggi X. Come è stato descritto in una nota precedente, dove questo metodo è stato applicato alla sequenza nucleotidica di geni, la sequenza di amminoacidi di una proteina globulare viene decomposta in « sequenze omopolipeptidiche difettive » contenenti residui di amminoacidi identici o simili e vacanze corrispondenti ai rimanenti residui di amminoacidi nella sequenza originale. Le sequenze così generate vengono trattate come reticoli lineari difettivi. Il metodo fornisce risultati molto promettenti nel riconoscimento di « omologie » e « analogie » fra sequenze di amminoacidi di proteine globulari di specie diverse e di una stessa specie. Inoltre consente di rivelare duplicazioni entro la struttura primaria di una proteina globulare. Ciò viene mostrato considerando come esempi il caso noto della Ferredoxina batterica del *Clostridium Pasterianum* ed il caso insospettato della Ferredoxina di spinaci.

Fourier analysis has been shown in a previous note [1] to be a powerful and most promising tool for revealing quasi-periodical patterns within the coding sequence of a gene. While this approach is being further developed and extended to a large number of genes, we wish to report some preliminary results obtained in a parallel application of the method to the analysis of the primary structures of globular proteins.

(*) Centro Interdisciplinare dell'Accademia Nazionale dei Lincei, Via della Lungara 10, 00165 Roma, Dipartimento di Chimica II Università di Roma (Tor Vergata), Roma.

(**) Istituto Chimico « G. Ciamician », Via Selmi 2, Università di Bologna, 40126 Bologna (Italy).

(***) Dipartimento di Chimica, Università di Roma (La Sapienza), P.le Aldo 00185 Moro, Roma.

(****) Centro di Studio per la Fisica delle Macromolecole del C.N.R., Via Selmi 2, 40126 Bologna (Italy).

(*****) Nella seduta del 23 giugno 1983.

Decomposition of the aminoacid sequence of a protein.

Just like the nucleotide sequence of a gene [1], the aminoacid sequence of a protein may in general be decomposed into 20 homopolypeptide sequences containing identical aminoacid residues and vacancies corresponding to the remaining missing aminoacid residues contained in the original aminoacid sequence.

For instance the initial aminoacid sequence of spinach ferredoxin [2] may be represented by the following set of one letter symbols [3]:

(1a) AAYKVTLVTPTGNV .....

It may be decomposed into the " defective homopolypeptide sequences ":

(1b) AAXXXXXXXXXXXXX .....

(1c) XXXXVXXVXXXXXV .....

(1d) XXXXXTXXTXTXXX .....

where the symbols A , V , T stands for Alanine, Valine, Threonine whereas the symbol X stand for a vacancy as above defined.

The aminoacid sequence of a globular protein may also be decomposed into " defective mixed polypeptides sequences " each containing vacancies X and aminoacid residues having similar physico-chemical properties, which may be expressed for instance by close values of their Van der Waals volumes, or by similar charged side chains.

Most of the aminoacid residues may accordingly be classified into several groups containing " quasi-isosteric " aminoacid residues. The most typical group contains the following hydrophobic aminoacid residues:

V (Val) , L (Leu) , I (Ile) , M (Met) , F (Phe).

A defective mixed polypeptide sequence containing only aminoacid residues of the above group and vacancies X may be generated from the aminoacid sequence 1a of spinach ferredoxin as follows:

(2) XXXXVXLVXXXXXV.....

Both defective homopolypeptide and mixed homopolypeptide sequences derived from the natural aminoacid sequence of a globular protein may be regarded as linear arrays of N elements each corresponding either to identical or to equivalent aminoacid residues or to vacancies X.

Sequence " homologies " and sequence " analogies ".

Similarities between the aminoacid sequences of polypeptide chains have been observed [4, 5, 6] by comparing either the aminoacid sequences of a given protein present in different species or the aminoacid sequences of different proteins present in the same species (usually called " protein families ").

It is operationally useful to distinguish between the above two kinds of sequence similarities by defining the former sequence " homologies " and the latter sequence " analogies ".

According to the above distinction, the aminoacid sequences of sperm-whale and human myoglobin are typical " homologous " sequences, whereas the aminoacid sequence of α and β globin chains of human haemoglobin are typical " analogous " sequences.

Going back to one of the old morphological distinctions between " homologies " and " analogies " introduced in Comparative Anatomy, two homologous proteins are related to each other like the feet of two mammals whereas two " analogous " proteins are related to each other like the hand and the foot of the same mammal.

Fourier Patterns of defective homopolypeptide and " mixed polypeptide sequences ".

The Fourier methods applied in the previous paper [1] to defective nucleotide sequences may be extended to both the above defined defective homopolypeptide and mixed polypeptide sequences. They may in fact be treated as defective linear lattices containing at lattice points either aminoacid residues or vacancies (X).

A function $D(x)$ may be calculated for the above defective linear lattices as a function of a linear coordinate x. It may be represented by the following Fourier series:

$$D(x) = \frac{1}{L} \sum_h |F(h)|^2 \exp\left(-2\pi i h \frac{x}{L}\right) \qquad h = 0, 1, 2, 3 \cdots . \tag{3}$$

The coefficients $|F(h)|^2$ are the squared moduli of the Fourier transform, which may be represented by:

$$F(h) = \sum_j \delta_j \exp\left(2\pi i h \frac{x_j}{L}\right) \qquad j = 1, 2, 3, \cdots, N \tag{4}$$

where $\delta_j = 1$ for an aminoacid residue and 0 for a vacancy X, x_j is the linear coordinate of a lattice point and L is the length of the unit cell of an imaginary super lattice, containing periodically translated defective sequences. The separation between two adjacent lattice points is assumed in all calculations to be unitary, while L is taken equal to 3.5 times the residue number N.

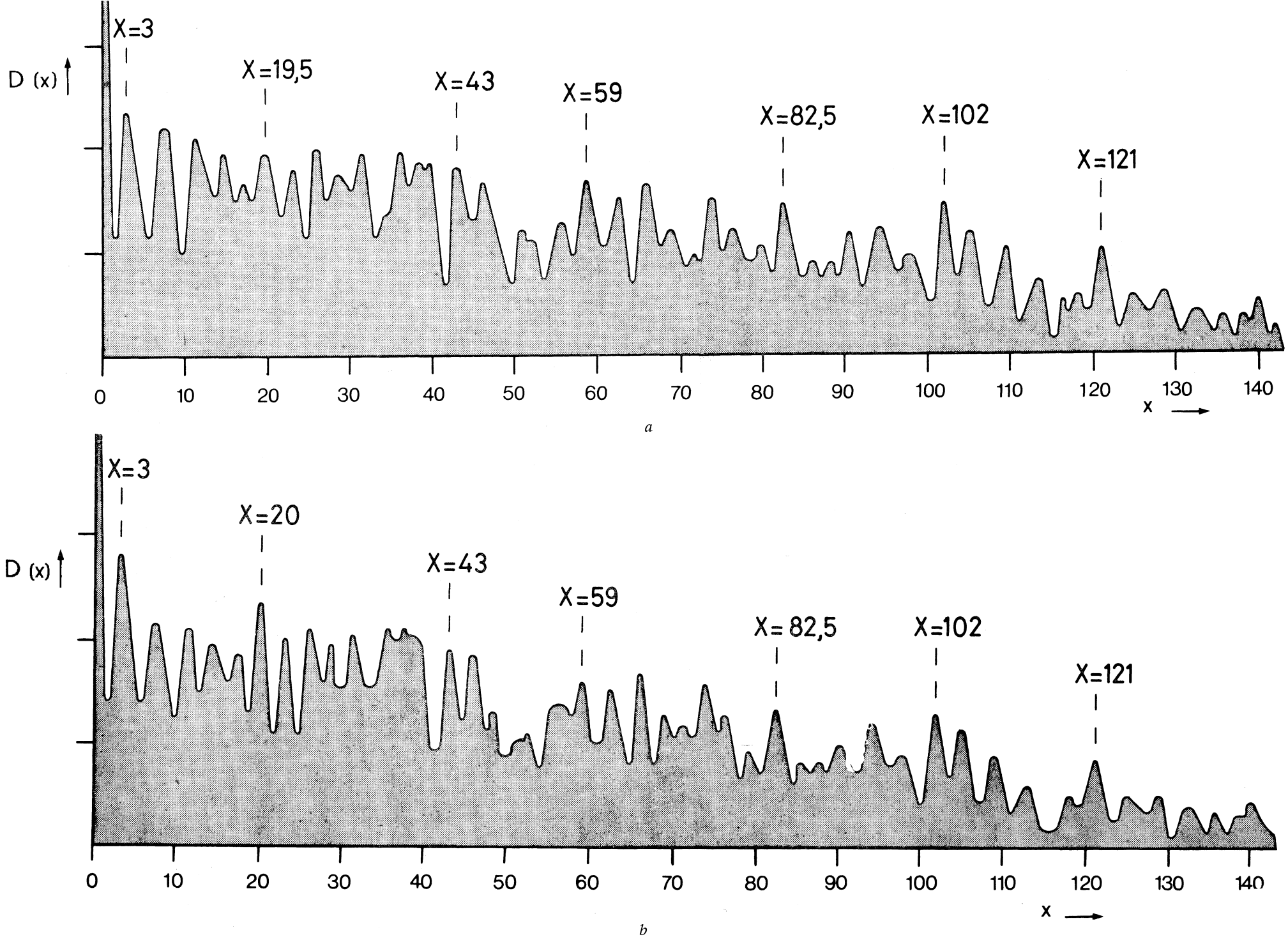
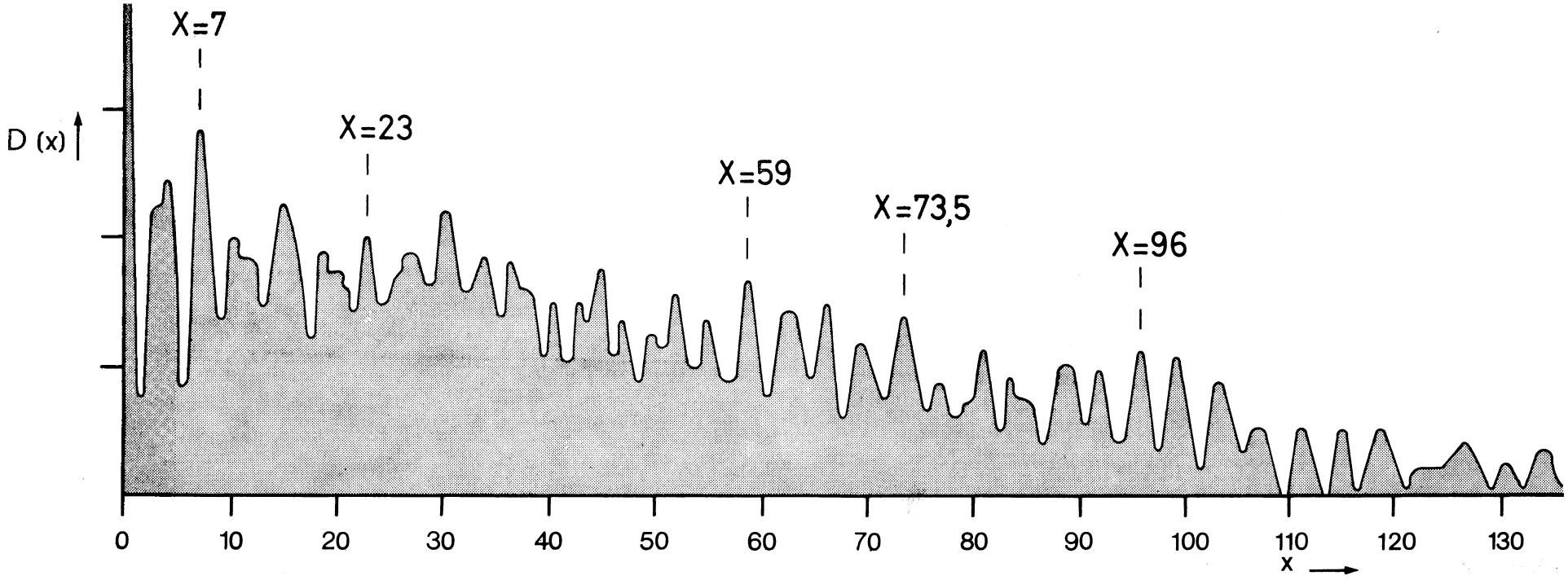
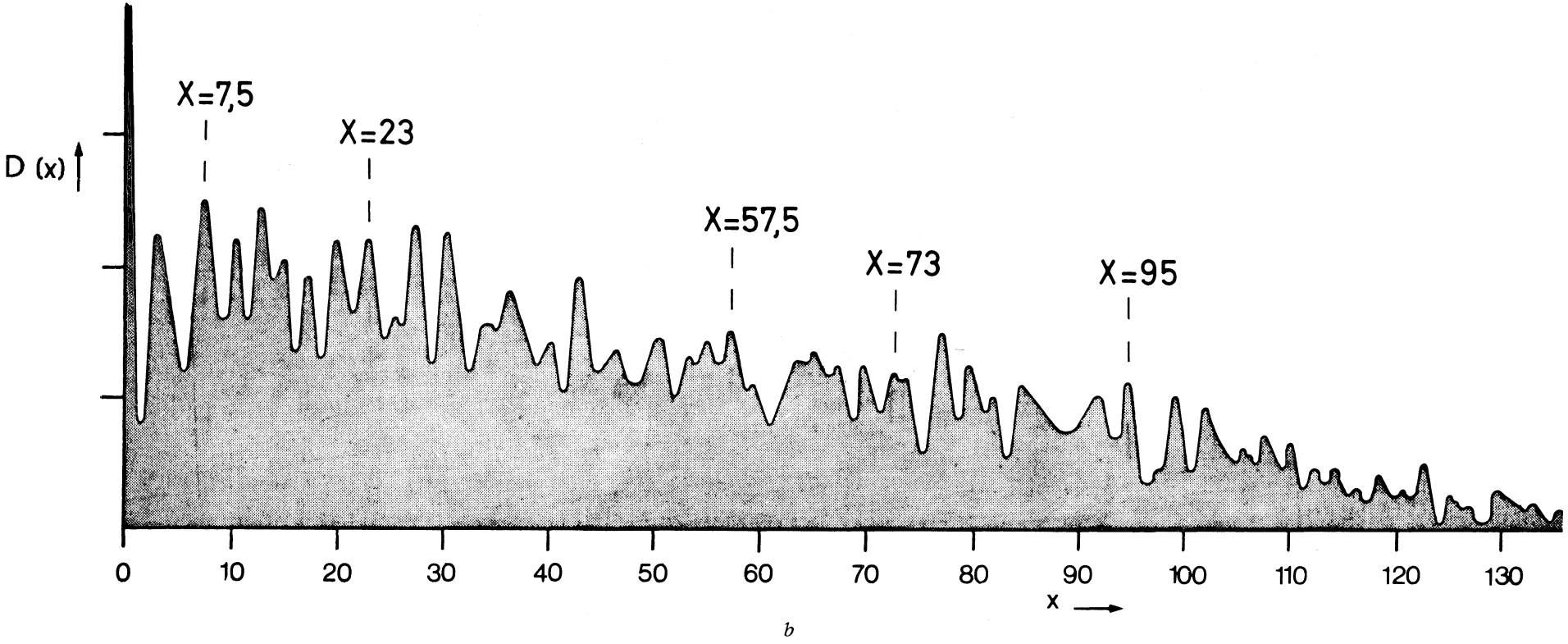


Fig. 1. – D (x) function calculated for human (a) and kangaroo (b) myoglobin considered as mixed polypeptide sequence containing hydrophobic residues (V , L , I , M , F) and vacancies X (unit cell lenght L = 536 residue number and h_{max} = 268).



a



b

Fig. 2. – D (x) functions calculated for α and β human globin (a, b respectively) considered as mixed polypeptide sequences containing hydrophobic residues and vacancies (unit cell length $L = 564$ and $h_{max} = 282$).

The above sum must be extended to the N lattice points corresponding to the N aminoacid residues contained in the natural aminoacid sequence of the protein from which the defective homopolypeptide and mixed polypeptide sequences have been derived. It should be noticed that for the latter the aminoacid residues belonging to a same group are treated as equivalent and therefore the coefficient δ_j appearing in 4) is set to unity for all the aminoacid residues of the group. The D (x) function is mathematically analogous to the " Patterson function " widely employed in the crystal structure analysis by X-ray diffraction methods [7]. A peak in the D (x) function corresponds to a distance d between two identical or equivalent aminoacid residues and the height of the peak is proportional to the number of times this distance occurs within the sequence. The resolution of the D (x) function depends upon the maximum h value of the terms of the Fourier series 3). All calculations, reported in this note, have been carried out with the Fourier series terminated at $h_{max} = L/2$. In order to reduce the series termination effects the $|F(h)|^2$ has been multiplied by a function $\exp(-Bh^2/4L^2)$ with $B = 10$.

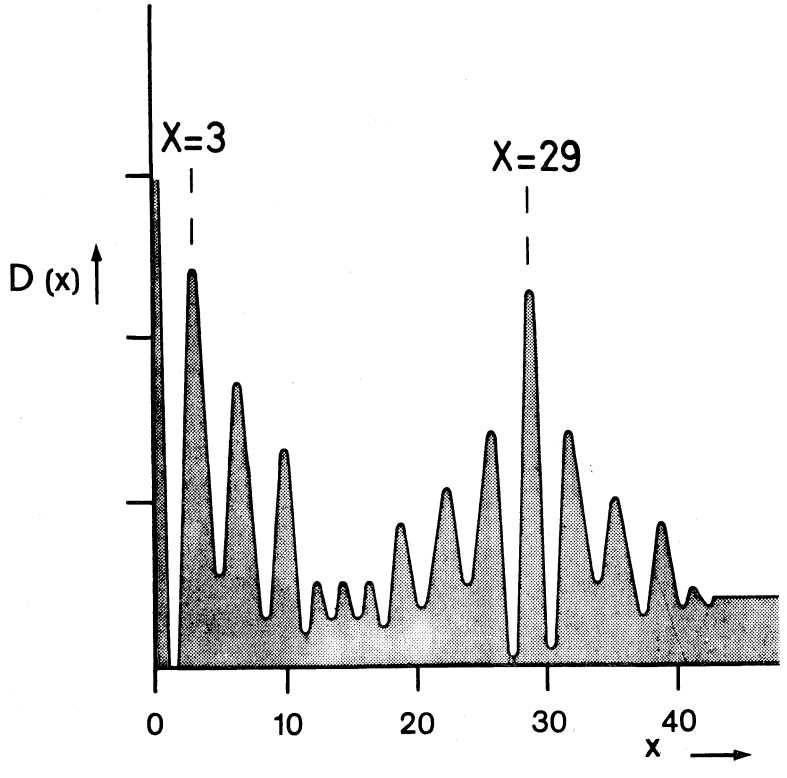


Fig. 3. - D (x) function calculated for ferredoxin of *Chlostridium Pasterianum* considered as homopolypeptide sequence containing only cysteine (C) and vacancies (X) (unit cell length $L = 192.5$ and $h_{max} = 94$).

Fourier pattern comparisons between "homologous" sequences and between "analogous" sequences

Mixed D (x) functions may be used for pattern recognition of similarities between homologous sequences and between analogous sequences. Mixed functions incorporate the observation that the most frequent aminoacid substitution observed when homologous proteins are compared is that occurring between "quasi-isosteric" aminoacid residues. The quasi-invariance of the average Van der Waals volume of a given protein, like cytochrome *c* or myoglobin [4], during biological evolution led, in fact, to the hypothesis that the aminoacid residues which are transmitted during biological evolution are those which do not alter the thermodynamic stability of the tertiary structure [6]. The above hypothesis has recently been confirmed by the close similarities found by X-ray diffraction between the tertiary structure of even evolutionary very distant molecules. The most striking examples are cytochrome *c* of rice and bonito [8] and lysozime of T4 bacteriophage and hen egg-white [9]. Figs. 1*a* and 1*b* show a comparison between the D (x) functions calculated for two evolutionary distant "homologous" myoglobin chains (human and kangaroo). Figs. 2*a* and 2*b* show a comparison between the D (x) functionscal culated for two "analogous" globin chains, namely the α and β globin of human haemoglobin. It may be observed that in both comparisons the pattern similarities between D (x) functions are very strong.

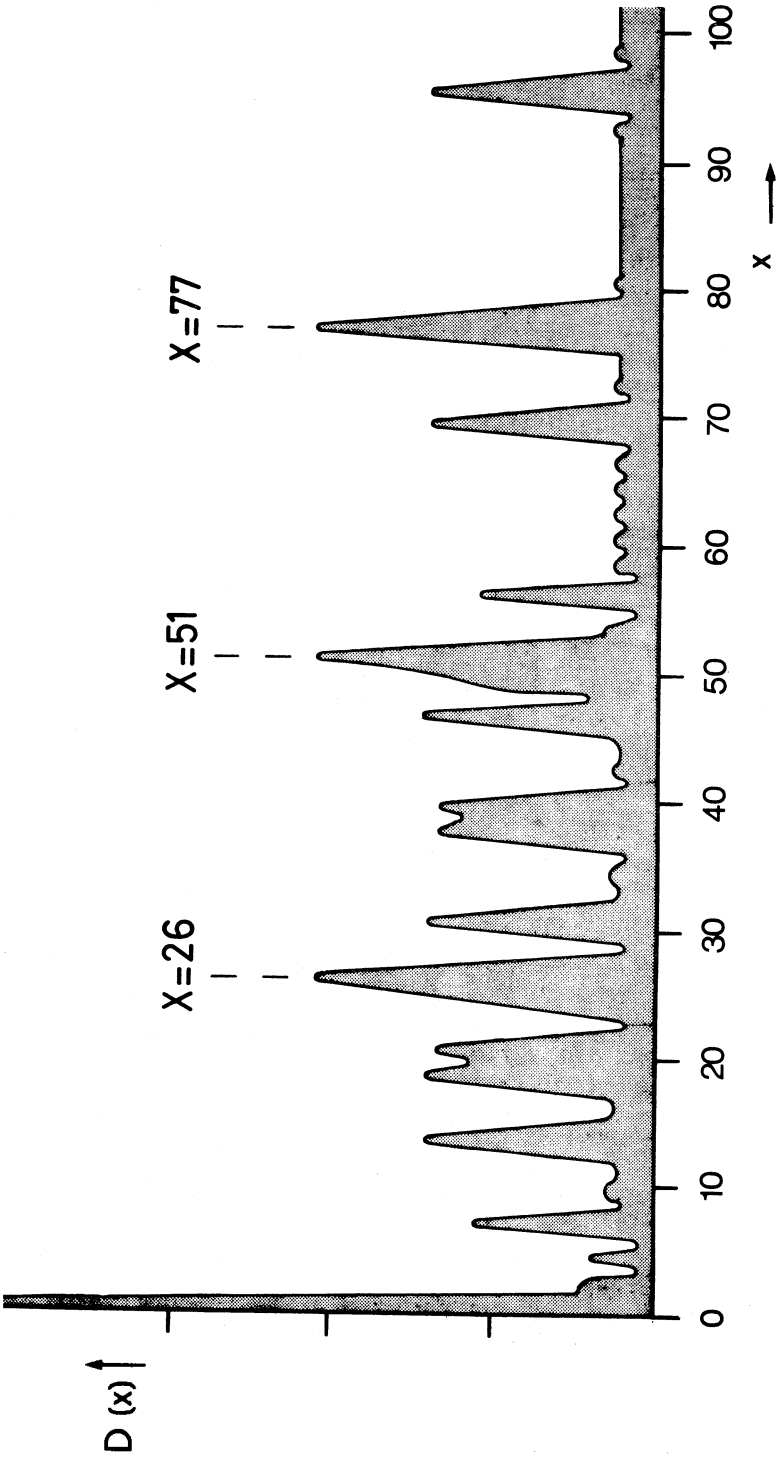
Detection of internal sequence duplications.

D (x) functions may be used as an objective method to detect the internal sequence duplication in a globular protein. This may be illustrated using as an example a bacterial protein, ferredoxin of *Chlostridium Pasterianum* [2]. It has been shown [3] that the aminoacid sequence of this protein may be self-aligned as follows:

AYKIXADSCVSCGACASECPVNAISQGDS

IFVIDADTCIDCGNCADVCPVGAPVQE

Fig. 3 shows the D (x) Pattern calculated for a defective homopolypeptide sequence containing only cysteine (C) (one of the most frequent aminoacid residues) and vacancies (X). It is characterized by a prominent peak corresponding to a repeated distance between every 29 identical aminoacid residues which is symmetrically flanked by lower peaks. Such a pattern strikingly reflects the internal sequence duplication. It should be stressed that the above pattern has been obtained without any assumption concerning aminoacid deletions.



a

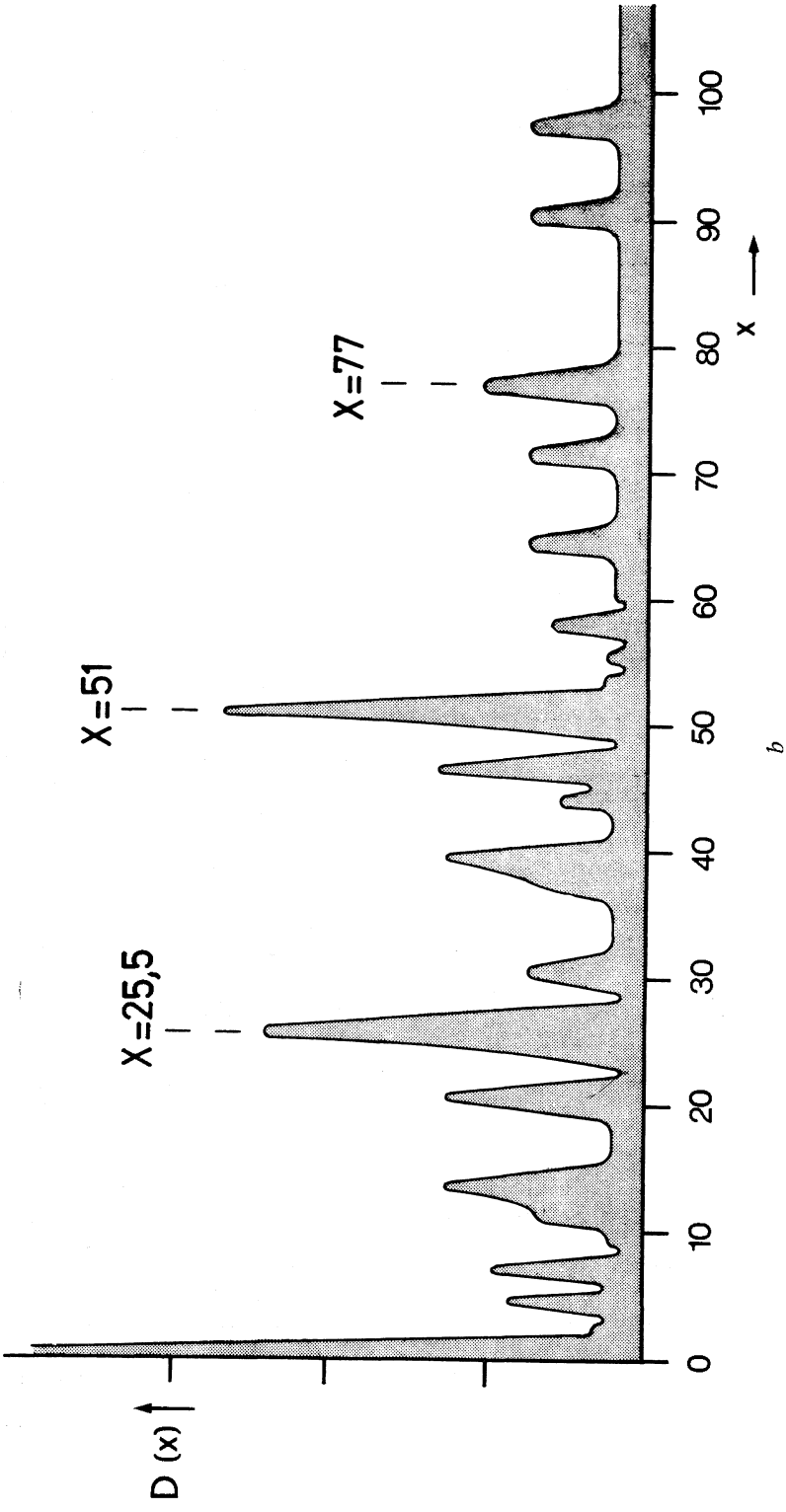


Fig. 4. – D (x) functions calculated for spinach ferredoxin considered as a defective homopolypeptide sequence containing only alanine /A) and vacancies (X) (a) and for a simulated defective homolypeptide sequence containing only alanine (A) and vacancies (X) (b) generated from the first half sequence (51 residues) of the native protein.

An even more interesting example of the power of D (*x*) functions for detecting sequence duplication of great interest in molecular evolution is shown in Fig. 4*a* by the pattern of the D (*x*) function calculated for a defective homopolypeptide sequence (containing identical very frequent alanine residues and vacancies X) derived from the aminoacid sequence of the spinach ferredoxin.

This protein contains 97 aminoacid residues (slightly less than twice the aminoacid residues of ferredoxin of *Chlostridium Pasterianum*). The D (*x*) pattern displays three prominent peaks at 26, 51, 77 spacing units, which may be interpreted by assuming that the aminoacid sequence of ferredoxin of spinach may be generated by duplication of an initial sequence of 26 aminoacid residues (homologous to the sequence of the ferredoxin of *Chlostridium Pasterianum*) followed by a further duplication of the above generated sequence.

Such an interpretation is convincingly supported by the pattern of the D (*x*) function calculated by simulating a defective homopolypeptide sequence generated from an initial sequence of the spinach ferredoxin containing 51 aminoacid residues (Fig. 4*b*).

It is hoped that further studies along the above outlined directions may contribute to a deeper understanding of molecular evolution and at the same time establish more precise correlations between primary and tertiary structures of globular proteins.

This work has been carried out with the financial support of C.N.R., Italy.

References

[1] A. M. Liquori, S. Ottani, A. Ripamonti and C. Sadun – « Rend. Acc. Naz. Lincei », in Press.

[2] G. D. Fasman (1976) – *Handbook of Biochemistry and molecular Biology*, « Proteins », Vol. III, ed. by R. C. Press, Cleveland Ohio.

[3] M. O. Dayoff (1972) – *Atlas of Protein Sequence and Structure*, « National Biomedical Research Foundation », Silver Spring Md.

[4] A. M. Liquori (1960) – *La diffrazione dei raggi X nello studio della costituzione molecolare di sostanze naturali*, Varenna 1958 « Ed. Accad. Naz. Lincei ».

[5] A. M. Liquori and C. Sadun (1976) – « Gazz. Chim. It. », *176*, 557.

[6] A. M. Liquori and C. Sadun (1980) – « Int. J. Biol. Macrom. ».

[7] A. M. Liquori (1981–83) in « Structural order in Polymers », p. 181, « J.U.P.A.C. Intern. Symp. », Florence Sept. Pergamon Press, Oxford and in « Conformation in Biology », p. 59, Ed. by R. Srinivasan and R. H. Sorma, Adenine Press, New York.

[8] M. Kakudo (1983) in « Conformation in Biology », p. 461. Ed. by R. Srinivasan and R. H. Sarma, Acenine Press, New York.

[9] R. Sarma (1983) in « Conformation in Biology », p. 69 ed. by R. Srinivasan and R. H. Sarma, Adenine Press, New York.