Математическая биология и биоинформатика. 2013. Т. 8. № 2. С. 708–715.
URL: http://www.matbio.org/2013/Teslyuk_8_708.pdf

========== ИНТЕЛЛЕКТУАЛЬНЫЙ АНАЛИЗ ДАННЫХ ==========

УДК: 519.22, 519.684

Применение метода главных компонент к анализу дифракционных изображений биомолекулярных объектов

Теслюк А.Б., Сенин Р.А., Ильин В.А.

Национальный исследовательский центр «Курчатовский институт», Москва 123182, Россия

Аннотация. В работе представлен алгоритм для быстрого детектирования изображений, содержащих дифракционную картину макромолекулярных объектов. Алгоритм основан на базе метода главных компонент, популярного алгоритма, используемого в различных задачах анализа многомерных данных, таких как классификация изображений, фильтрация шумов, индексация видео и др. В данной работе мы показываем эффективность применения алгоритма для анализа рентгенограмм различных макромолекулярных структур, содержащих коллаген, полученных на Курчатовском центре синхротронного излучения.

Ключевые слова: *анализ изображений, кластеризация данных, метод главных компонент, методы снижения размерности, дифракция биомолекулярных объектов, анализ дифракционных данных, классификация дифракционных изображений.*

ВВЕДЕНИЕ

Строящиеся и уже построенные импульсные источники рентгеновского диапазона – лазеры на свободных электронах – обещают получение кардинально новой информации о структуре белковых молекул. Такие источники, называемые также четвертым поколением синхротронных источников, уже построены в Стэнфорде (США) [1], строятся в Цукубе (Япония) [2] и Гамбурге (ФРГ) [3]. На запущенном лазере уже получены первые результаты по характеризации кластеров белковых молекул и нанокристаллов белков [4].

Экспериментальные данные в таких экспериментах поступают с очень высокой скоростью – порядка 5 Gb в секунду, или, с учетом необходимых дополнительных операций, поток экспериментальных данных поток в сутки оценивается примерно в 200 Тб [5]. Из них полезными являются, по оценкам, около 4 Тб, а остальное – неудачные импульсы лазера, а также промахи мимо мишеней либо мишени, содержащие только исходный раствор, без кристаллов интересующих исследователей белков.

Очевидно, что в таких условиях важной задачей является селекция экспериментальных данных, и маркировка их для последующей обработки.

Задача представляет интерес, т. к. объем данных весьма велик – в приведенной статье [4] использовано порядка 1,8 млн. удачных дифракционных картин. Система сортировки должна эффективно отбраковывать неудачные снимки, поступающие со скоростью порядка 160 дифракционных картин в секунду в случае «медленного» лазера в Стэнфорде, либо до 27 тыс. импульсов на европейском лазере в ДЭЗИ.

В качестве первой ступени снижения числа некорректных данных на Европейском лазере предлагается использовать систему veto – запрета записи данных с заведомо некачественным пучком [6]. Однако это не снимает проблему неудачных образцов.

Решением может стать фильтрация экспериментальных данных «на лету», но при этом необходимы алгоритмы и методы для такой фильтрации, позволяющие оперативно оценить непригодность данных.

Одним из популярных методов анализа данных высокой размерности, к которым можно отнести дифракционные изображения, является метод главных компонент [7]. Метод главных компонент является одним из мощных и универсальных средств анализа, который, не отбрасывая конкретные признаки, позволяет учитывать лишь наиболее значимые комбинации их значений.

При его использовании в задаче классификации изображений, каждое изображение разлагается на линейную комбинацию собственных векторов, которые называются главными компонентами. В этом случае главные компоненты могут быть представлены в виде изображений.

Метод главных компонент активно применяется для задач анализа дифракционных данных. Так в работе [8] предлагается метод для определения структурных свойств карбонных нанотрубок с помощью метода главных компонент. В работе [9] авторы предлагают метод главных компонент для анализа геометрических свойств ряда химических соединений, а в работе [10] приводится пример удачного применения метода для анализа спектра лекарственных соединений.

Мы предлагаем применять этот метод для определения наличия дифракционной картины в изображениях рентгеновского рассеяния. Среди преимуществ метода главных компонент для анализа потока данных рентгеновского рассеяния следует отметить:

- Высокую эффективность метода для задачи фильтрации шумов в изображениях и поиска наиболее характерных особенностей в потоке изображений.
- Простоту реализации. В основе метода лежат линейные преобразования входных и сингулярное разложение матрицы, что эффективно реализовано в большинстве математических библиотек и прикладных пакетов.
- Возможность параллельной обработки данных и простого внедрения в высокопроизводительных вычислительных комплексах.

Математические библиотеки, такие как Intel MKL (Math Kernel Library), используемые на параллельных кластерах, содержат реализации необходимых методов и алгоритмов, которые лежат в основе метода главных компонент. Поэтому модификация программ для обработки и фильтрации изображений потребует минимальных изменений, а именно перекомпиляции с параллельными библиотеками для применения на суперкомпьютерах.

ОПИСАНИЕ МЕТОДА

В данном разделе мы приводим математическое описание метода кластеризации многомерных данных с помощью метода главных компонент. Допустим, у нас имеется k изображений, каждое из которых можно представить в виде матрицы пикселей n на m или в виде $n \times m$-мерного вектора $x_i = [x_1, x_2, ..., x_{mn}]$, в котором все столбцы матрицы пикселей выстроены в один столбец. Таким образом, размерность исходного пространства данных равна $n \times m$. В случае анализа дифракционных изображений размерность данных может быть достаточно велика, и для дальнейшего анализа удобно преобразовать данные, чтобы понизить размерность, потеряв, при этом, как можно меньше информации.

Карл Пирсон [7] предложил для понижения размерности перейти от базиса исходных признаков (в случае изображений – пикселей) к базису из собственных

векторов ковариационной матрицы данных, упорядоченных по убыванию собственных чисел. Собственные векторы ковариационной матрицы называются главными компонентами, а порожденный ими базис – базисом главных компонент. Одно из свойств базиса главных компонент – максимальная дисперсия проекций данных на подпространство, порожденное первыми l главными компонентами, по сравнению с любыми другими подпространствами размерности l. Например, при $l = 2$ базис, построенный на первых двух главных компонентах, дает нам плоскость, проекции на которую исходных данных показывают максимальное разнообразие.

Для поиска базиса главных компонент обычно применяют метод сингулярного разложения матрицы. Расположим k $n \times m$-мерных векторов, соответствующим анализируемым изображениям в виде столбцов матрицы $\mathbf{X}$ размерности $(n \times m,\ k)$. Предварительно данные необходимо центрировать: $x_i' = x_i - \overline{x}_i$. Нам необходимо найти собственные векторы ковариационной матрицы $\mathbf{A} = \frac{1}{k}\mathbf{X}^T\mathbf{X}$:

$$(\mathbf{A} - \lambda_k \mathbf{E})\nu_k = 0\,. \qquad (1)$$

Для этого, воспользуемся сингулярным разложением матрицы $\mathbf{X}$ на произведение двух ортогональных и одной диагональной матрицы:

$$\mathbf{X} = \mathbf{U}\Sigma\mathbf{V}^T\,, \qquad (2)$$

где $\mathbf{U}$ и $\mathbf{V}$ - ортогональные матрицы, а S – диагональная. Нетрудно видеть, подставив определение матрицы A в (2): $\mathbf{A} = \frac{1}{k^2}\mathbf{U}^T(\Sigma^T\Sigma)\mathbf{U}$, что столбцы ортогональной матрицы $\mathbf{U}$ и есть главные компоненты, а элементы диагональной матрицы $\Sigma^T\Sigma$ – её главные значения. Таким образом, задача поиска базиса главных компонент сводится к задаче сингулярного разложения матрицы центрированных исходных данных.

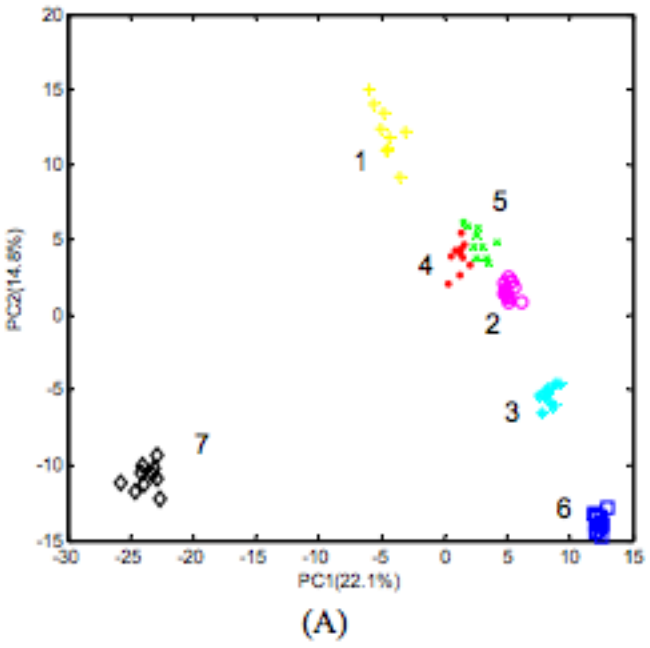

Рис. 1. Пример кластеризации рентгеновских снимков (из [5]).

Получив базис главных компонент, мы можем уменьшить размерность пространства данных до плоскости. Затем с помощью метода k средних редуцированное пространство входных данных делится на два кластера: изображения, содержащие дифракционную картину, и на изображения, не содержащие таковую. Подобный метод широко применяется для различных задач классификации, например

авторы [8] аналогичным методом эффективно определяли молекулярный состав вещества по рентгеновским изображениям. Пример кластеризации рентгеновских снимков различных веществ приведен на рисунке 1.

ФИЛЬТРАЦИЯ ДИФРАКЦИОННЫХ ИЗОБРАЖЕНИЙ

Первым нашим шагом была проверка возможности фильтрации изображений, где отсутствует дифракционная картина. Для этого в качестве обучающей выборки был взят массив дифракционных картин, соответствующих различным объектам, полученный ранее в Курчатовском Центре Синхротронного излучения. Отобраны были, также, и настроечные изображения без объектов. Примеры изображений, содержащих и не содержащих дифракционную картину, приведены на рисунках 2 и 3.

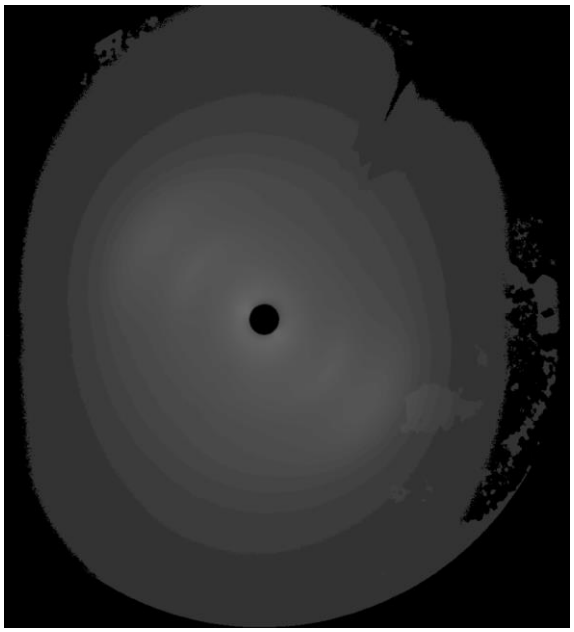

Рис. 2. Пример дифракционной картины на молекулах синтетического волокна.

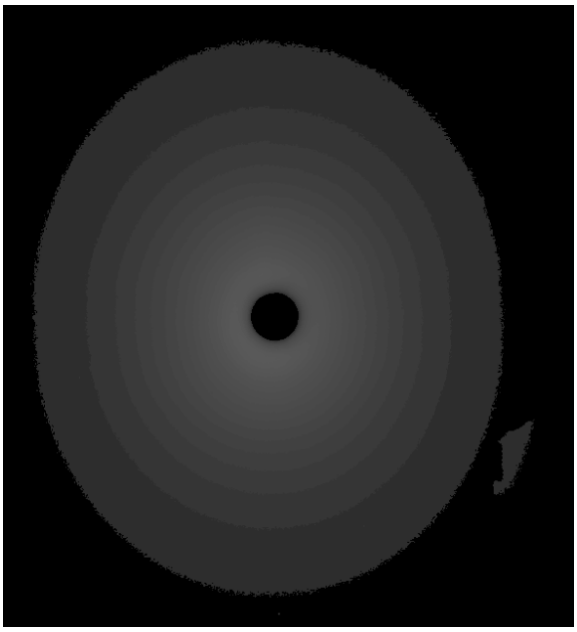

Рис. 3. Пример изображения, не содержащего дифракционную картину. В части изображения можно видеть шумы и дефекты считывающей матрицы.

В качестве обучающей выборки было использовано около 50 изображений каждого класса, затем был произведен поиск базиса главных компонент и построен график, соответствующий точкам исходных изображений в пространстве первых двух компонент. Этот график приведен на рисунке 4. На графике видно, что точки, соответствующие изображениям, не содержащим дифракционную картину, очень хорошо локализуются в пространстве первых двух главных компонент. Можно легко выделить область в этом пространстве, и, сравнивая координаты неизвестных изображений с границами этой области, быстро определять те изображения, где дифракционная картина отсутствует.

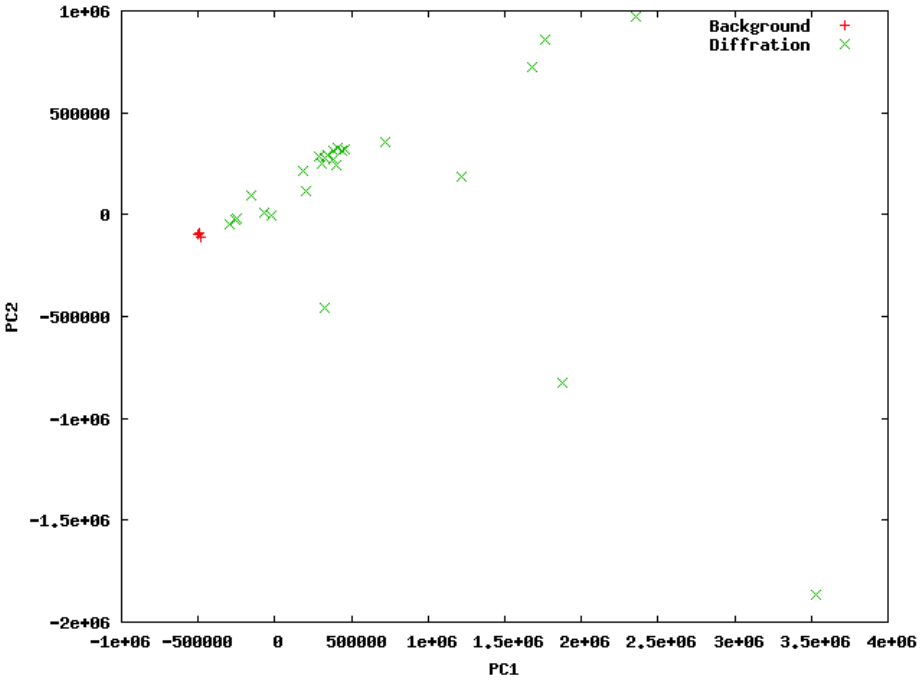

Рис. 4. Два класса изображений в пространстве главных компонент.

ПРОВЕРКА ВОЗМОЖНОСТИ КЛАССИФИКАЦИИ ИЗОБРАЖЕНИЙ, СОДЕРЖАЩИХ ДИФРАКЦИОННУЮ КАРТИНУ РАЗЛИЧНЫХ ОБЪЕКТОВ

Следующий шаг состоял в том, чтобы, имея дифракционные изображения различных объектов, понять, можно ли быстро определять, какому объекту соответствует неизвестное изображение. Для этого мы использовали массив данных, содержащих дифракционные картины трех сходных объектов, содержащих соединительную ткань и белок коллаген – изображения клеток кожи лосося, роговицы оболочки глаза кролика и клеток амниотической оболочки кролика. Примеры дифракционных изображений приведены на рис 5–7.

Рис. 5. Дифракция на клетках кожи лосося.

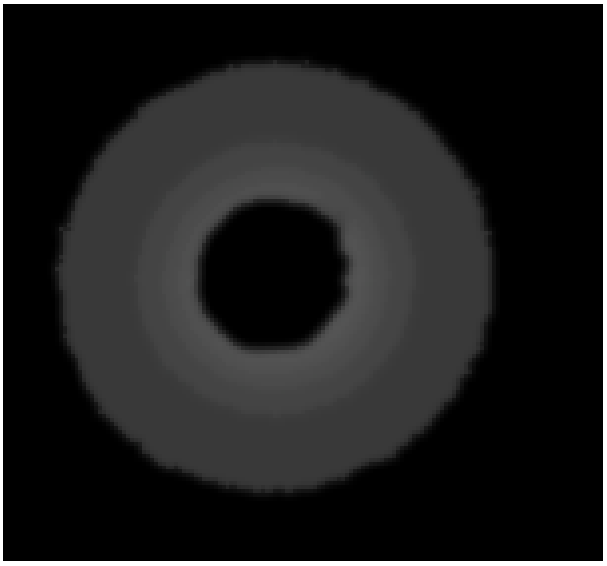

Рис. 6. Дифракция на роговице глаза кролика.

Рис. 7. Дифракция на амниотической оболочке кролика.

Все эти объекты содержат родственные белки – коллагены [12, 13], отличающиеся своей структурой [14]. Использовалось около 20 изображений каждого класса. Картина расположения обучающих изображений в пространстве главных компонент представлена на рисунке 8. Из рисунка видно, что изображения, соответствующие дифракции на коже лосося, хорошо отличимы от остальных изображений. В то же время точки, соответствующие рентгенограммам амниотической оболочки и роговицы глаза, достаточно сильно пересекаются, что, видимо, связано с наличием в них одних и тех же белков.

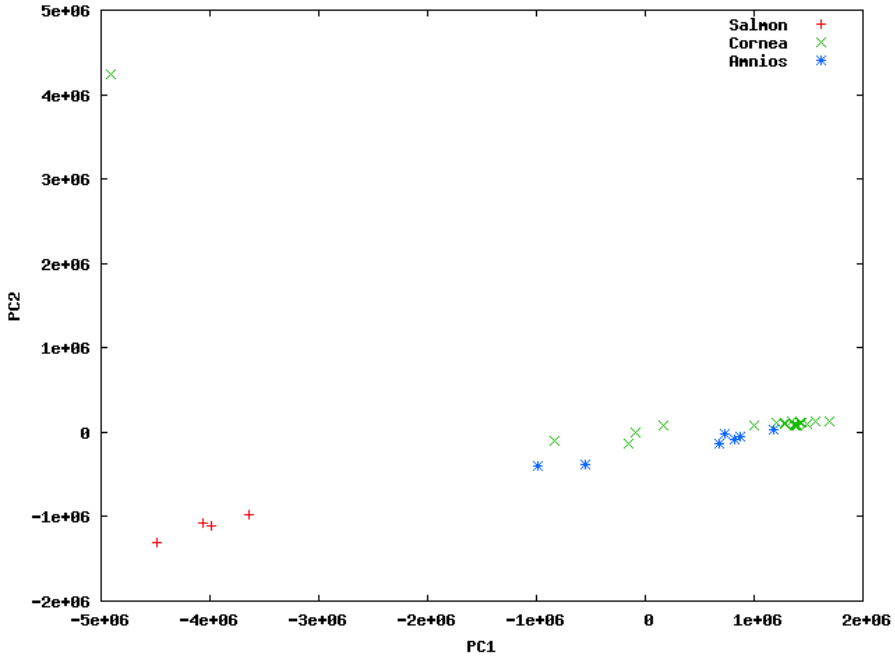

Рис. 8. Три класса изображений в пространстве главных компонент.

Стоит отметить что изображения, которыми мы пользовались для тестирования метода, значительно отличаются от тех, что планируется получить в проекте XFEL, где речь идет о дифракции на отдельных молекулах. В дальнейшем мы планируем провести обучение системы на дифракционных изображениях, которые получены в результате моделирования взаимодействия макромолекул с электромагнитным излучением, что позволит нам настроить классификационные правила более точно.

ЗАКЛЮЧЕНИЕ

В настоящей работе описан алгоритм для фильтрации изображений, не содержащих дифракционную картину, а также показана его эффективность для анализа реальных данных, полученных в Курчатовском центре синхротронного излучения. Нами показана 100-процентная точность фильтрации изображений, не содержащих дифракционную картину. Мы полагаем, что разработанный алгоритм может быть успешно применен для фильтрации изображений, получаемых в экспериментах по исследованию пространственной структуры макромолекул с помощью лазеров на свободных электронах.

Кроме того, нам удалось показать возможность классификации дифракционных изображений ряда сходных структур, содержащих коллаген. Такой подход классификации может быть использован для построения самообучающейся системы, которая может автоматически классифицировать дифракционные изображения, получающиеся в ходе эксперимента, по степени сходства с изображениями из обучающего банка данных. Кроме задачи классификации может решаться задача индексации дифракционных картин и построения базы данных с возможностью поиска изображений, сходных с неизвестным образцом.

В дальнейшем планируется доработка алгоритма классификации, которая позволит более точно классифицировать дифракционные изображения и определять исследуемые молекулярные объекты. Для этого планируется использовать в качестве дескрипторов изображений значения угловых автокорреляционных функций интенсивностей рентгенограмм. В ряде работ показана связь таких функций с пространственной структурой молекул [14–16], что совместно с применением метода главных компонент может позволить более точно определять дифракционные изображения, соответствующие различным молекулам. Подобный подход позволит эффективно выделять дифракционные изображения интересующих макромолекул из общего потока экспериментальных данных, что, в свою очередь, повысит эффективность экспериментов по исследованию пространственной структуры сложных биологических молекулярных объектов.

Авторы выражают благодарность за предоставленные дифракционные данные коллективу станции "Дикси" Курчатовского источника синхротронного излучения, А.В. Забелину, А.Ю. Грузинову и руководителю станции Альвине Андреевне Вазиной, зав. лаб. Института теоретической и экспериментальной биофизики РАН, г. Пущино.

СПИСОК ЛИТЕРАТУРЫ

1. DiMauro L.F., Arthur J., Berrah N., Bozek J., Galayda J.N., Hastings J. Progress report on the LCLS XFEL at SLAC. *J. Phys.: Conf. Ser*. 2007. V. 88. P. 012058.
2. RIKEN. First X-ray lasing of SACLA: Next-generation facility up and running with powerful new X-ray laser. *ScienceDaily*. 2011. URL: http://www.sciencedaily.com/releases/2011/06/110613053815.htm (дата обращения 16.12.2013).
3. European X-Ray Free-Electron Laser tunnel construction completed. *Phys.org*: web-based science, research and technology news service. 2012. URL: http://phys.org/news/2012-06-european-x-ray-free-electron-laser-tunnel.html (дата обращения 16.12.2013).
4. Chapman H.N., Fromme P., Barty A., White T.A., Kirian R.A., Aquila A., Hunter M.S., Schulz J., DePonte D.P., Weierstall U. et al. Femtosecond X-ray protein nanocrystallography. *Letters to Nature*. 2011. V. 470. P. 73–78.
5. Youngman C. Data acquisition and Controls. *XFEL Users' Meeting*. 2012. URL: http://www.xfel.eu/sites/site_xfel-

gmbh/content/e63594/e65073/e126274/e134393/3Youngman_DataAcquisitionandControls_eng.pdf (дата обращения 16.12.2013).

6. Mancuso A.P., Aquila A., Borchers G., Giewekemeyer K., Reimers N. Scientific Instrument Single Particles, Clusters, and Biomolecules (SPB) *Technical Design Report*. 2013. URL: https://docs.xfel.eu/alfresco/d/a/workspace/SpacesStore/497 (дата обращения 16.12.2013).
7. Pearson K. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine.* 1901. V. 2. P. 559–572.
8. Oddershede J., Nielsen K., Stahl K. Using X-ray powder diffraction and principal component analysis to determine structural properties for bulk samples of multiwall carbon nanotubes. *Zeitschrift für Kristallographie.* 2007. V. 222. P. 186–192.
9. Bosco J.P. da Silva, Ramos M.N. Principal Component Analysis of Molecular Geometries of Cis- and Trans-C2H2X2. *J. Braz. Chem. Soc.* 2004. V. 15. N. 1. P. 43–49.
10. Li W., Zhong Y., Yu D., Qu D., Sun B., Li M., Liu J. Application of principal component analysis for identification of drugs packed in anthropomorphic phantom. *ARPN Journal of Engineering and Applied Sciences.* 2012. V. 7. N. 7. P. 915–921.
11. Yata M., Yoshida C., Fujisawa S., Mizuta S., Yoshinaka R. Identification and characterization of molecular species of collagen in fish skin. *Journal of Food Science.* 2011. V. 66. P. 247–251.
12. Cintron C., Hong B.S., Covington H.I., Macarak E.J. Heterogeneity of collagens in rabbit cornea: type III collagen. *Invest. Ophthalmol. Vis. Sci*. 1988. V. 29. N. 5. P. 767–775.
13. Bornstein P., Sage H. Structurally distinct collagen types. *Annual review of biochemistry*. 1980. V. 49. P. 957–1003.
14. Altarelli M., Kurta R., Vartanyants I.A. X-ray cross-correlation analysis and local symmetries of disordered systems: General theory *Phys. Rev. B.* 2010. V. 82. № 10. P. 104207.
15. Kurta R.P., Ostrovskii B.I., Singer A., Gorobtsov O.Y., Shabalin A., Dzhigaev D., Yefanov O.M., Zozulya A.V., Sprung M., Vartanyants I.A. X-ray cross-correlation analysis of liquid crystal membranes in the vicinity of the hexatic-smectic phase transition. *Phys. Rev. E*. 2013. V. 88. P. 044501.
16. Kurta R.P., Dronyak R., Altarelli M., Weckert E., Vartanyants I.A. Solution of the phase problem for coherent scattering from a disordered system of identical particles. *New J. Phys*. 2013. V. 15. P. 013059.

Материал поступил в редакцию 06.12.2013, опубликован 30.12.2013.