Available online at www.sciencedirect.com
ScienceDirect
Journal of Structural Biology 157 (2007) 47–55
Journal of Structural Biology
www.elsevier.com/locate/yjsbi

SPARX, a new environment for Cryo-EM image processing

Michael Hohn [a], Grant Tang [b], Grant Goodyear [c], P.R. Baldwin [c], Zhong Huang [c], Pawel A. Penczek [c], Chao Yang [e], Robert M. Glaeser [d], Paul D. Adams [a], Steven J. Ludtke [b,*]

[a] *Lawrence Berkeley National Laboratory, Physical Bioscience Division, Berkeley, CA 94720, USA*
[b] *National Center for Macromolecular Imaging, Verna and Marrs McLean Department of Biochemistry and Molecular Biology, Baylor College of Medicine, Houston, TX 77030, USA*
[c] *The University of Texas, Houston Medical Center, Department of Biochemistry and Molecular Biology, Houston, TX 77030, USA*
[d] *The University of California, Department of Molecular and Cell Biology, Physical Biosciences Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA*
[e] *Lawrence Berkeley National Laboratory, Computational Research Division, Berkeley, CA 94720, USA*

Received 9 May 2006; received in revised form 1 July 2006; accepted 7 July 2006
Available online 16 July 2006

Abstract

SPARX (single particle analysis for resolution extension) is a new image processing environment with a particular emphasis on transmission electron microscopy (TEM) structure determination. It includes a graphical user interface that provides a complete graphical programming environment with a novel data/process-flow infrastructure, an extensive library of Python scripts that perform specific TEM-related computational tasks, and a core library of fundamental C++ image processing functions. In addition, SPARX relies on the EMAN2 library and *cctbx*, the open-source computational crystallography library from PHENIX. The design of the system is such that future inclusion of other image processing libraries is a straightforward task. The SPARX infrastructure intelligently handles retention of intermediate values, even those inside programming structures such as loops and function calls. SPARX and all dependencies are free for academic use and available with complete source.

Keywords: CryoEM; TEM; Single particle analysis; 3-D reconstruction; Image processing

1. Introduction

Numerous excellent software packages are available for the TEM community, including SPIDER (Frank et al., 1996), IMAGIC (van Heel et al., 1996), BSOFT (Heymann, 2001), FREALIGN (Grigorieff, 1998), EM (Hegerl, 1996), IMIRS (Liang et al., 2002), SUPRIM (Schroeter and Bretaudiere, 1996), IMOD (Kremer et al., 1996), PHOELIX (Carragher et al., 1996), PFT (Baker and Cheng, 1996), the MRC reconstruction tools (Crowther et al., 1996) and Xmipp (Sorzano et al., 2004). Each of these packages has its own strengths and weaknesses, and although the general thrust of the methodologies is the same, each has its own set of particularly well developed methods used to achieve a final reconstruction. However, because of varying file formats, different conventions for Euler angles and the parameterization of the contrast transfer function (CTF) of the instrument, parameters of the image formation process, etc., moving between packages in order to take advantage of their respective strengths can be an exceedingly difficult and time consuming process.

Over the last two decades, single particle reconstruction has gone from a technique that initially was capable of achieving structures in the 20–30 Å range to a versatile mainstream tool frequently producing structures at subnanometer resolution, and in a few pioneering projects, approaching 4–5 Å resolution. One of the requirements to achieve high resolution is to limit the electron dose

* Corresponding author.
E-mail address: sludtke@bcm.edu (S.J. Ludtke).

1047-8477/$ - see front matter © 2006 Elsevier Inc. All rights reserved.
doi:10.1016/j.jsb.2006.07.003

delivered to the specimen in order to minimize the impact of radiation damage. Reduction of the dose causes a corresponding decrease in the signal-to-noise ratio in the image. Therefore, the techniques required to achieve high resolution structures must become ever more sophisticated in order to cope with very noisy data and at the same time to deliver highly accurate alignment parameters.

Numerous computational methods are used in a single particle reconstruction, and for each there are various choices of competing algorithms. For example, it is possible to compute a 3-D reconstruction from projection data using algebraic real-space methods often implemented as iterative algorithms, such as ART (Gordon et al., 1970) or SIRT (Lakshminarayanan and Lent, 1979), filtered back-projection methods, or direct Fourier inversion methods implemented in Fourier space (for detailed review see (Vainshtein and Penczek, 2006)). For 2-D alignment of noisy particles to low-noise projections, there are exhaustive search algorithms, iterative algorithms which separate the rotational alignment from the translational alignment, methods based on moments of the two images, and many others. Beyond this, there are choices between different methodologies for the overall reconstruction process, some of which use so-called 'inverse' methods, and avoid the reconstruction step entirely, as discussed below. In no case can a single algorithm be selected and make a claim to be 'the best' algorithm for all situations. Different algorithms react differently to variations in overall signal-to-noise ratio, shape of the object and CTF parameters. To discover the optimal methods for a particular reconstruction, it is often necessary to try many possible variants of the available algorithms. Incorporation of all or most possibilities into a single environment would dramatically ease this process of determining what works best in a particular experiment, and, in fact, may lead to a better understanding of why specific algorithms perform better in specific situations.

An additional difficulty arising in virtually all of the aforementioned software is the difficulty in organizing datasets which often consist of tens if not hundreds of thousands of particles drawn from tens to thousands of micrographs or CCD frames. Frequently, obtaining a final reconstruction involves gradually paring the data down to include only the highest quality micrographs, and/or only the highest quality particles from each micrograph. This process is a substantial data organization task, generally handled manually through careful organization of tiered directories containing hundreds of files each. This rapidly becomes untenable as reconstructions grow from a few thousand particles from tens of images to hundreds of thousands of particles from thousands of images. Development of an automatic data tracking system integrated with the reconstruction software has become an important task in the further development of single particle processing.

The goal in creating SPARX is to provide a uniform environment for end-users, in which to combine elements of different structure determination strategies, and develop their own approaches, without being forced to acquaint themselves with all available software suites. High-level strategies can be easily modified using a graphical programming approach, which does not require detailed knowledge of algorithms made available by the designers of the system. Issues such as file format conversion and Euler angle conventions are dealt with as automatically and transparently as possible. SPARX offers a core of robust image processing capabilities and an easy to use programming environment.

X-ray crystallography and cryo-EM single particle reconstruction are powerful techniques for macromolecular structure determination at intermediate to high resolutions. One of the goals of SPARX is to provide an integrated computational environment for both methods. This is of increasing importance as they are now often used concurrently. Crystallographic reconstructions of components of a macromolecular assembly can be readily combined with lower resolution structures of intact complexes solved by cryo-EM. In the future, as cryo-EM methods move to higher resolution it will be possible to take advantage of existing crystallographic tools for electron density interpretation. Conversely, crystallographic methods can benefit from the use of cryo-EM envelopes for phasing.

2. SPARX design

In recent years, a wide range of scientific disciplines have adopted the Python scripting language (http://www.python.org) in conjunction with low-level C++ code to address the need for flexibility without sacrificing performance. As discussed in Section 3, as an easy to learn object-oriented scripting language, Python permits very rapid development and debugging of new routines. Thus, SPARX makes use of C++ for compute-intensive code, while Python is used to implement complex higher level image processing tasks. The link between C++ and Python is generated using the Boost Python Library (Abrahams and Grosse-Kunstleve, 2003). This same approach has been used successfully in the PHENIX software suite for automated crystallographic structure determination (Adams et al., 2002, 2004).

The overall design of SPARX is diagrammed in Fig. 1. Users can interact with SPARX in three different ways: (i) through a graphical programming interface, which requires no formal programming background, (ii) through use of pre-written scripts from a command shell, (iii) through a text-based customized Python interpreter. The SPARX C++ core library integrates three components: a set of algorithms written specifically by SPARX developers, the core image processing library from the EMAN2 package (see the companion piece in this issue) and *cctbx* (Grosse-Kunstleve et al., 2002) from the PHENIX project, providing tools for manipulating crystallographic data and molecular models. Expanding SPARX by adding simple links to other image processing packages is a simple process. To demonstrate this, a partial wrapper for SPIDER (Frank et al., 1996) is provided, making use of the file format and