Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

Rapid shape determination of tissue transglutaminase using high-throughput computing

Donna Lammie,[a]* James Osborne,[b] Daniel Aeschlimann[c] and Timothy J. Wess[a]

[a]School of Optometry and Vision Sciences, Cardiff University, Maindy Road, Cardiff CF24 4LU, Wales, [b]Information Services, Cardiff University, 50 Park Place, Cardiff CF10 3AT, Wales, and [c]School of Dentistry, Cardiff University, Heath Park, Cardiff CF14 4XY, Wales

Correspondence e-mail: lammied@cf.ac.uk

Received 10 April 2007
Accepted 5 July 2007

Small-angle X-ray scattering can be used to determine the molecular shape of macromolecules in solution which are otherwise refractory to conventional high-resolution studies. *DAMMIN* and *GASBOR* are applications that utilize *ab initio* methods to build models of proteins using simulated annealing; both *DAMMIN* and *GASBOR* have to be run numerous times on the same input data to generate the most likely protein shape. *Condor* is a specialized workload-management system for PC computation-intensive tasks. Using *Condor*, *DAMMIN* and *GASBOR* can be run a number of times simultaneously on the same input data, allowing the shape of proteins to be determined in a fraction of the time it would have taken to have run *DAMMIN* and *GASBOR* sequentially. The main advantage of this approach is that it allows quicker data processing; therefore, results are obtained promptly and without undue delay. Tissue transglutaminase is a multidomain enzyme that catalyses the formation of isopeptide bonds between polypeptide chains. This reaction requires the enzyme to undergo a series of conformational changes that are not well understood in order to allow the sequential interaction with the two substrate proteins and their subsequent release when cross-linked. *Condor* was applied to determine the solution shape of tissue transglutaminase in a rapid fashion. Eventually, the next step will be to move towards online analysis at synchrotron sources by developing a graphical user interface that will enable remote access, allowing users to submit jobs to *Condor* whilst at synchrotrons.

1. Small-angle X-ray scattering

Small-angle X-ray scattering (SAXS) is used to investigate the structure of macromolecules on the nanometre length scale and can be employed to characterize the molecular shape of proteins in solution. The scattering pattern provides information about the size and shape of proteins and also their interactions. The structural detail acquired by SAXS can be related to information obtained at different levels of architecture, thus allowing the structure of complex biological systems and the basis of how they are assembled to be understood. This would also provide the potential for the structure–function relationship of proteins to be studied. The density probability profile in solution, when combined with complementary information such as the three-dimensional atomic resolution structure obtained from X-ray crystallography or nuclear magnetic resonance (NMR), makes SAXS an extremely useful technique (Grossmann, 2002).

2. *DAMMIN* and *GASBOR*

The size and shape of molecules in solution can be extracted from the scattering pattern using a series of computer algorithms. *DAMMIN* (Svergun, 1999) is a computer program that uses an *ab initio* method to build models of the protein shape by simulated annealing using a single-phase dummy-atoms model. The program *GASBOR* (Svergun *et al.*, 2001) is used to analyse the data and uses similar parameters to *DAMMIN*; however, instead of the dummy-atom model, an ensemble of dummy residues are used to form a chain-compatible model. Given that *DAMMIN* and *GASBOR* utilize *ab initio* methods to build models of proteins, they are typically run a number of times on the

same input data. The output files from these programs are input files to another series of programs, *DAMAVER* (Volkov & Svergun, 2003), which aligns the models from *DAMMIN* and *GASBOR* and produces an averaged filtered model. The greater the number of repetitions of *DAMMIN* and *GASBOR*, the more accurate the final protein shape.

3. *Condor*: high-throughput computing

Condor is a specialized workload-management system for the execution of computation-intensive tasks (Litzkow *et al.*, 1988) using a network of computer workstations. Further information can be found on the Condor project's website at http://www.cs.wisc.edu/condor, where the *Condor* software and complete documentation are freely available. The *Condor* pool at Cardiff University has an average of around 1400 execute nodes providing 800 gigaflops of computing power that is available on demand for users. Users submit their jobs to *Condor*, which places them into a queue, decides where and when to run individual tasks based upon a policy, monitors their progress and informs the user upon completion. Through the use of *Condor*, *DAMMIN* and *GASBOR* can be run multiple times simultaneously for one or more proteins, thus allowing the work to be completed rapidly and efficiently. For example, using the runs conducted on tissue transglutaminase, the total time for 20 repeat runs would have been approximately 12 h on one PC. Using *Condor*, 20 repeat runs were performed in approximately 36 min, representing a significant performance gain in terms of accessibility.

DAMMIN and *GASBOR* can be used in interactive mode, which requires the user to input a number of parameters before processing the output from *GNOM*, an indirect Fourier transform program (Semenyuk & Svergun, 1991), and generating a model of the protein that can be visualized. Typically, when running *DAMMIN* and *GASBOR* in interactive mode, the user is prompted to answer questions such as symmetry and expected particle shape; if such answers are not known, then the default responses are accepted. The user has to input the name of the *GNOM* file, the log file (used to log any errors) and the project identifier (used to name the output file). In addition to the interactive mode, both programs can now be run in batch mode, where the user only has to input the answers to the most important parameters as listed in Konarev *et al.* (2006).

A similar approach was applied in this case, where a submit script generator (SSG) was developed to assist the user in running *DAMMIN* and *GASBOR* using the *Condor* toolkit. More information about the SSG can be found on the Cardiff University website at http://www.cf.ac.uk/optom/research/condor.html. The SSG was developed to submit jobs specifically to *Condor*, but with minimal adaptations could be used to submit to other workload-management systems such as, for example, *GLOBUS* or *PBS*. The SSG asks the user only once for the necessary information to prepare and submit multiple jobs to *Condor*, thereby reducing the time taken to submit and process multiple proteins. If 20 simulations are to be run, the SSG looks in the specified input directory for *GNOM* files and generates 20 answer files containing the name of the *GNOM* file, a log file in the format `file0.log` to `file19.log` and a project identifier in the

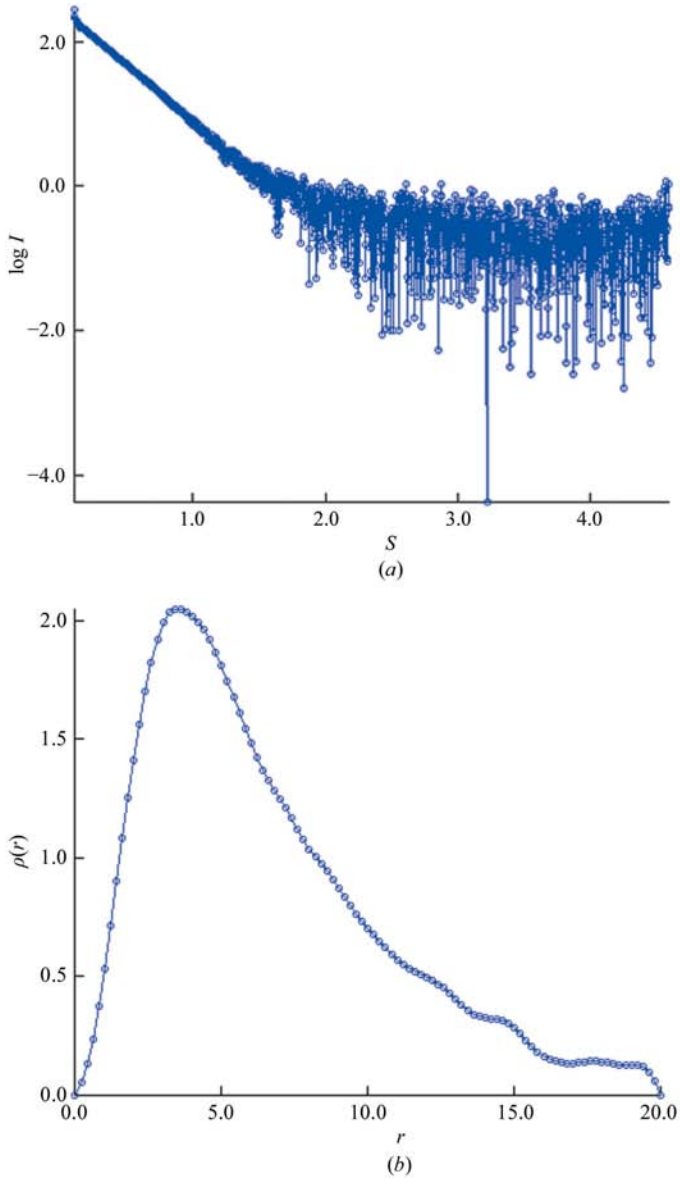

Figure 1
(*a*) The one-dimensional profile of the experimental data after the buffer background has been subtracted. (*b*) The particle distance distribution function $\rho(r)$ produced by *GNOM*. (*c*) The most probable shape of transglutaminase 2 in solution in the absence of Ca^{2+} and GTP obtained from the average of 20 independent simulations produced from *DAMMIN* (blue). The crystal structure of TG2 with bound GDP has been superimposed (red).

format `file0` to `file19`. The SSG only prompts the user to input those parameters where the user wishes to consider an answer other than the default input and the SSG can be adjusted if further user input is required. The SSG then produces 20 batch files, each of which calls *DAMMIN* or *GASBOR* with one of the 20 answer files. The SSG then builds the *Condor* submit script itself which, when submitted, instructs *Condor* to transfer a copy of the *DAMMIN* or *GASBOR* binary, a *GNOM* file, an answer file and a batch file to an execute node and instructs *Condor* to transfer the output files back to the user's workstation when the run is complete.

Recently, *Condor* has assisted in running multiple simulations of *DAMMIN* and *GASBOR* to determine the shape of fibrillin-1, an extracellular matrix protein involved in tissue elasticity (Baldock *et al.*, 2006), and in the shape determination of a subfragment of the tropoelastin molecule (Dyksterhuis *et al.*, 2007).

4. Transglutaminase

Transglutaminases are a family of enzymes that are capable of introducing isopeptide bonds in or between polypeptide chains by catalyzing a Ca^{2+}-dependent transfer reaction between the γ-carboxamide group of a peptide-bound glutamine residue and a primary amine, most commonly the ε-amino group of a lysine residue (Folk & Finlayson, 1977). The action of these enzymes consequently results in the formation of covalently cross-linked, often insoluble supramolecular structures and has a well established role in tissue homeostasis in many biological systems (Aeschlimann & Thomazy, 2000; Lorand & Graham, 2003). Cross-linking in proteins is important in providing stability to any ordered conformations with which they are compatible (Creighton, 1997); for example, cross-linking mediated by tissue transglutaminase (transglutaminase 2) plays a role in extracellular matrix (ECM) stabilization and thereby promotes strengthening of the cell-matrix adhesion apparatus (Stephens *et al.*, 2004).

5. Data collection and analysis

As our example, data on human recombinant tissue transglutaminase (2 mg ml^{-1}) were collected at station X33 of the European Molecular Biology Laboratory (EMBL) at the Deutsches Elektronen Synchrotron (Hamburg, Germany). The two-dimensional data was converted into one-dimensional linear profiles using in-house software at station X33. Values obtained for buffer solution (20 m*M* Tris–HCl pH 7.2, 100 m*M* NaCl) without protein were subtracted from the data using *PRIMUS* (Konarev *et al.*, 2003); the corrected profile is shown in Fig. 1(*a*). *GNOM* was used to estimate the particle distance distribution function, $\rho(r)$, as shown in Fig. 1(*b*), from the experimental scattering data. The results are in good agreement with small-angle neutron scattering data (Mariani *et al.*, 2000). The output files produced by *GNOM* were entered into *DAMMIN* and *GASBOR* and 20 simulations of each were conducted, which was facilitated by the use of *Condor*. The average filtered shape of tissue transglutaminase is shown in Fig. 1(*c*). The crystal structure of TG2 with bound GDP has been superimposed on this envelope for comparison and reveals a good fit. Any slight discrepancies with the fit could possibly be a consequence of the molecular envelope being produced from TG2 only and the crystal structure consisting of TG2 and bound GDP.

6. Conclusions

This paper has described the simultaneous execution of numerous *DAMMIN* and *GASBOR* runs by distributing the jobs to a network of PCs using *Condor*, which has permitted a substantial acceleration of the processing of results. Thus, the main advantage of *Condor* is that it allows rapid turnover of experimental data since results can be processed quickly and efficiently. The distribution of *DAMMIN* and *GASBOR* runs using *Condor* would be beneficial to many researchers that use these programs to perform similar scattering experiments or structure-determination studies. In order to progress with fast data analysis, the next step will be to parallelize *DAMAVER*, thus accelerating the averaging process.

Increased throughput design at new SAXS beamlines and automation at existing SAXS beamlines should be matched by software solutions that allow rapid online as close to real-time analysis as possible. Currently being developed is an experimental design with a graphical user interface that will enable remote access, allowing users to submit jobs to *Condor* during data acquisition at synchrotrons. The ability to guide the experimental procedures during data collection will allow the user to assess and evaluate the results immediately after collection. This will permit the user to adapt experiments instantly, if necessary, thus allowing the user to make the most efficient use of their time at the synchrotron.

We would like to thank the staff at station X33 at EMBL, Hamburg, Germany for their assistance. We are grateful to M. Langley and P. Aeschlimann for technical assistance, and to the Bardhan Research and Education Trust of Rotherham (BRET) for funding.

References

Aeschlimann, D. & Thomazy, V. (2000). *Connect. Tissue Res.* **41**, 1–27.
Baldock, C., Siegler, V., Bax, D. V., Cain, S. A., Mellody, K. T., Marson, A., Haston, J. L., Berry, R., Wang, M. C., Grossmann, J. G., Roessle, M., Kielty, C. M. & Wess, T. J. (2006). *Proc. Natl Acad. Sci. USA*, **103**, 11922–11927.
Creighton, T. E. (1997). *Proteins: Structures and Molecular Properties*, 2nd ed. New York: W. H. Freeman & Co.
Dyksterhuis, L. B., Baldock, C., Lammie, D., Wess, T. J. & Weiss, A. S. (2007). *Matrix Biol.* **26**, 125–135.
Folk, J. E. & Finlayson, J. S. (1977). *Adv. Protein Chem.* **31**, 1–133.
Grossmann, J. G. (2002). *Scattering and Inverse Scattering in Pure and Applied Science*, edited by R. Pike & P. Sabatier, pp. 1123–1139. New York: Academic Press.
Konarev, P. V., Petoukhov, M. V., Volkov, V. V. & Svergun, D. I. (2006). *J. Appl. Cryst.* **39**, 277–286.
Konarev, P. V., Volkov, V. V., Sokolova, A. V., Koch, M. H. J. & Svergun, D. I. (2003). *J. Appl. Cryst.* **36**, 1277–1282.
Litzkow, M., Livny, M. & Mutka, M. (1988). *Proceedings of the Eighth IEEE International Conference on Distributed Computing Systems*, pp. 104–111. New York: IEEE.
Lorand, L. & Graham, R. M. (2003). *Nature Rev. Mol. Cell Biol.* **14**, 140–156.
Mariani, P., Carsughi, F., Spinozzi, F., Romanzetti, S., Meier, G., Casadio, R. & Bergamini, C. M. (2000). *Biophys. J.* **78**, 3240–3251.
Semenyuk, A. V. & Svergun, D. I. (1991). *J. Appl. Cryst.* **24**, 537–540.
Stephens, P., Grenard, P., Aeschlimann, P., Langley, M., Blain, E., Errington, R., Kipling, D., Thomas, D. & Aeschlimann, D. (2004). *J. Cell Sci.* **117**, 3389–3403.
Svergun, D. I. (1999). *Biophys. J.* **76**, 2879–2886.
Svergun, D. I., Petoukhov, M. V. & Koch, M. H. J. (2001). *Biophys. J.* **80**, 2946–2953.
Volkov, V. V. & Svergun, D. I. (2003). *J. Appl. Cryst.* **36**, 860–864.