electronic reprint

Journal of
Applied Crystallography
ISSN 0021-8898

GraphEnt: a maximum-entropy program with graphics capabilities

Nicholas M. Glykos and Michael Kokkinidis

Journal of
Applied Crystallography
ISSN 0021-8898

Received 8 December 1999
Accepted 16 March 2000

GraphEnt: a maximum-entropy program with graphics capabilities

Nicholas M. Glykos[a]* and Michael Kokkinidis[a,b]

[a]IMBB, FORTH, PO Box 1527, 71110 Heraklion, Crete, Greece, and [b]Department of Biology, University of Crete, PO Box 2208, 71409 Heraklion, Crete, Greece. Correspondence e-mail: glykos@crystal2.imbb.forth.gr

A maximum-entropy formalism aimed at the production of a 'maximally noncommittal' map is a standard method in fields of science like radio-astronomy, but a rare exception in both X-ray crystallography and electron microscopy (or crystallography). This is rather unfortunate, given the wealth of information that a maximum-entropy map can reveal, especially when the map itself is the end product (for example, low-resolution electron or potential density maps, Patterson functions, deformation maps). The program *GraphEnt* attempts to automate the procedure of calculating maximum-entropy maps, with emphasis on the calculation of difference Patterson functions for macromolecular crystallographic problems, while providing a useful graphical output of the current stage of the calculation.

1. Introduction

The principle of maximum entropy has been a consistent source of controversy ever since its introduction by Jaynes (1957). The proponents of the method argue that the maximum-entropy ('MaxEnt') principle is the only consistent method of statistical inference, but this statement appears to be open to debate (see, for example, Skilling, 1984; Uffink, 1995; and references therein). Given the controversy that still surrounds the theoretical foundations of the method, it is rather surprising how its practical applications (which cover most fields of scientific analysis) have outrun the theoretical studies, to the point of judging the value of the principle on the basis of the quality of the results obtained from it. Although this pragmatic approach entails the danger of misusing the method, it is probably fair to say that every data analysis problem is at its heart pragmatic, in the sense that the analysis is performed with the expectation of extracting some justifiable (by the data) conclusions. If a method has repeatedly been shown to produce results that appear to be superior to those produced by other methods, then the temptation to postpone its theoretical justification for the future is both strong and understandable.

Most applications of the maximum-entropy method in the field of macromolecular crystallography have focused on the phase determination/extension/refinement problems, mainly through the pioneering work by Bricogne (1984, 1997) and co-workers (also reviewed by Gilmore, 1996). On the other hand, very little progress has been made with respect to the practical day-to-day application of the maximum-entropy principle for the calculation of crystallographic maps. We believe that with the currently available computing power, the problem is not so much the cost (in terms of CPU time) of calculating such a map, as it is the absence of freely available software for automatically performing the calculation. The program *GraphEnt*, described herein, is the result of an attempt to produce software capable of performing the unsupervised calculation of a maximum-entropy map consistent with a set of crystallographic observations.

2. Algorithms, implementation and program specification

2.1. Algorithms

GraphEnt maximizes the configurational entropy of the map subject to the constraint that the final map is consistent with the observed data using a modification of the algorithms of Gull & Daniell (1978) and Collins (1982). Although this algorithm is neither the most efficient nor the most stable, it is relatively easy to code and it leads, at least in the case of Patterson syntheses (where the phases are fixed and known), to essentially the same results as other, more complex algorithms (see, for example, Bricogne, 1984; Smith & Grandy, 1985).

2.2. Software environment

Programming language and operating systems: the program is written in ANSI C and, although it has been developed in a Unix[1] environment, is expected to be portable to any computer system with an ANSI-compliant C compiler [with the provision that some of the additional features of the program (§2.4) may not be available in a non-Unix environment].

Overlay structure: none.

Subroutine libraries accessed: the minimum requirement for successful compilation and linking of *GraphEnt* is the availability of the freely distributed FFTW library (Frigo & Johnson, 1998; obtainable *via* http://www.fftw.org/). If X-windows-based graphics support is required, the freely distributed PGPLOT library is also needed (http://astro.caltech.edu/~tjp/pgplot/). For those users wishing to have direct support of the binary *CCP*4 reflection and map files

[1] Trademarks: Unix is a registered trademark of Unix System Laboratories, Inc. Postscript is a registered trademark of Adobe Systems, Inc. Silicon Graphics, Origin 200, Indigo2, O_2, Indy and Irix are trademarks of Silicon Graphics, Inc. (Mountain View, CA). R5000, R4600 and R10000 are trademarks of MIPS Technologies, Inc. DEC, DEC Alpha 1200 and OSF are trademarks of Compaq, Inc.

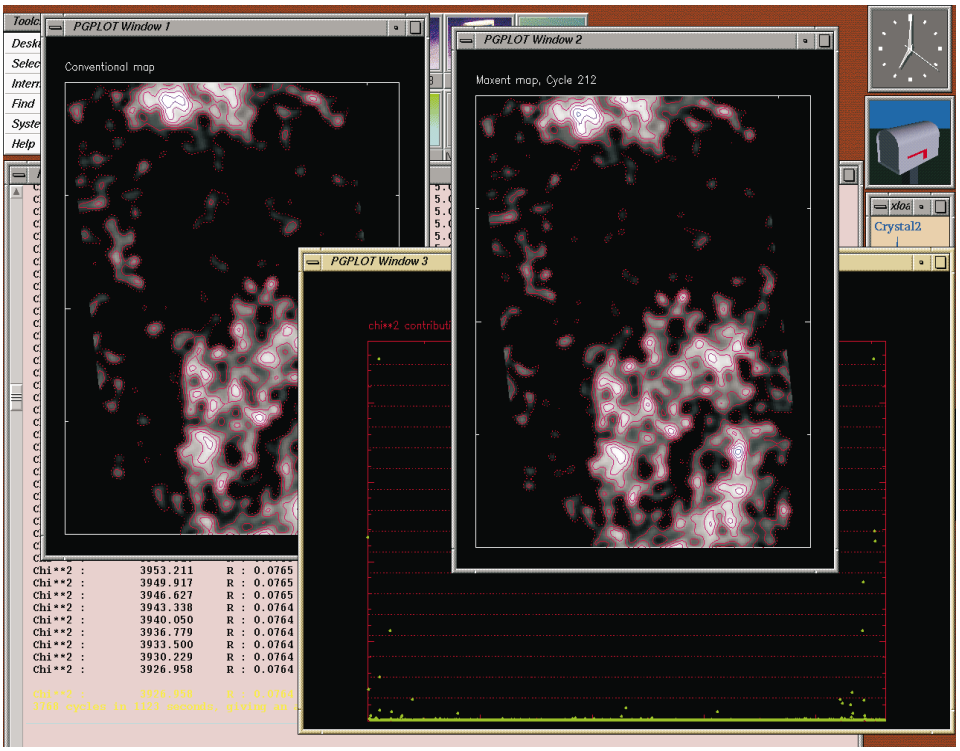

Figure 1
A snapshot of *GraphEnt* in action.

(Collaborative Computational Project, Number 4, 1994), the corresponding library must be available at compilation time.

2.3. Hardware environment

Computers and installation: Silicon Graphics computers O_2 R5000 Irix 6.3, O_2 R10000 Irix 6.3, Indigo2 R10000 Irix 6.2, Origin 200 R10000 Irix 6.4 and Indy R4600 Irix 6.2; DEC Alpha OSF computers DEC Alpha server 1200, OSF1, V4.0. The stand-alone executable can be located in any suitable directory.

Minimum number of bits per byte: 32.

Minimum size of physical memory required: at least six times the size of the map in bytes.

2.4. Program specification

Restrictions on the complexity of the calculation: the calculation is always performed in space group *P*1, and consequently, there are no space-group-specific restrictions. The type of crystallographic syntheses that *GraphEnt* can automatically recognize and perform are the following: Patterson syntheses [defined by h, k, l, F, $\sigma(F)$], difference Patterson syntheses [h, k, l, F_1, $\sigma(F_1)$, F_2, $\sigma(F_2)$], phased but unweighted syntheses [h, k, l, F, $\sigma(F)$, φ] and figure-of-merit (FOM) weighted syntheses [defined by h, k, l, F, $\sigma(F)$, φ, FOM].

Data formats: the input to the program is either a free-format ASCII file containing a list of reflections, or a binary *CCP*4 (`.mtz`) reflection file. The supported output map formats are either ASCII formatted files or binary *CCP*4 map files.

Typical run times: these depend greatly on the size of the map, the type of calculation and the quality of the input data. For example, a 262 144 (= 128 × 64 × 32) pixels $mF_o\exp i\varphi_{best}$) map corresponding to a reasonably accurate (by macromolecular standards) 3.8 Å data set was calculated in less than 8 min of CPU time on a DEC Alpha 1200, while a 524 288 (= 128 × 128 × 32) pixels difference Patterson map for a weakly substituted derivative (which makes the calculation easier) took only 46 s on the same machine. On the other hand, a 2 Å ($2mF_o - DF_c$)$\exp(i\varphi_c$) synthesis with 3 072 000 (= 160 × 160 × 120) pixels took ~40 min of CPU time.

Number of lines: 5537 for the source code, 3197 for the raw *LATEX* document.

Test status: several difference Patterson functions for three different crystal forms have been calculated, both in projection and in three dimensions. The program has also been tested with a medium-resolution single isomorphous replacement (SIR) phased protein map, and with an 8 Å resolution cryoelectron-microscopic reconstruction of the potential density projection of a large multiprotein complex.

Additional features: *GraphEnt* uses the PGPLOT graphics library to plot (using contours and/or grayscale representations) a user-defined section from both the conventional and the maximum-entropy maps. The plot of the maximum-entropy map is updated as the calculation proceeds, allowing the user to identify its most persistent features. Fig. 1 shows an image captured from the screen of a workstation performing a *GraphEnt* calculation. In the case of an isomorphous difference Patterson calculation, the program also draws the corresponding normal probability plot (Howell & Smith, 1992) which can be used to select suspect data points.

Table 1
Crystal data and statistical information on the test data set.

Space group	$P2_1$
Unit-cell parameters (Å, °)	a = 35.9
	b = 28.9
	c = 63.6
	β = 105.5
Resolution (Å)	20.4–1.8
Total number of observations	34875
Number of unique reflections	8712
Overall R_{sym}	0.044
Overall completeness (%)	73.1
Overall $I/\sigma(I)$	20.4
R_{sym} for 1.86–1.80 Å	0.096
Completeness for 1.86–1.80 Å (%)	71.2
$I/\sigma(I)$ for 1.86–1.80 Å	9.8
χ^2 test on anomalous differences	1.63
R_{sym} when Bijvoet pairs not merged	0.039
Completeness for Bijvoet pairs (%)	61.5

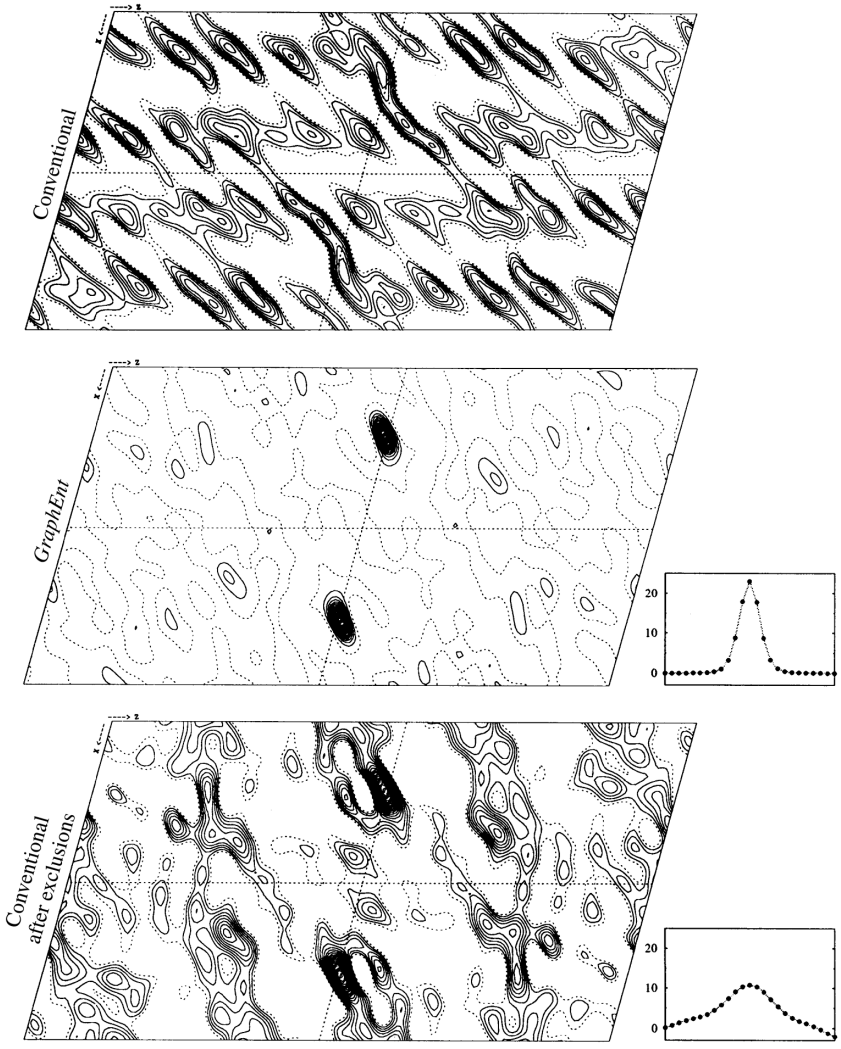

Figure 2
Comparison of the Harker (v = 1/2) sections from two conventional anomalous Patterson functions and a *GraphEnt* anomalous Patterson function of a myoglobin crystal (see text for details). All three Harker sections are contoured with the first (dashed) contour at the mean density, and then at intervals of 0.5 of the r.m.s. deviation of the densities of the whole map. The two insets show the distribution of normalized density [in units of $(\rho - \langle\rho\rangle)/\sigma(\rho)$] through the major peaks (of the respective maps) and in a direction parallel to the longest axes of the peaks. This figure was prepared with the programs *PLUTO* and *PLTDEV* from the *CCP*4 suite of programs (Collaborative Computational Project, Number 4, 1994) and with the program *XMGR*, available *via* http://plasma-gate.weizmann.ac.il/Xmgr/.

2.5. Documentation

Extensive documentation is available with the distribution, in the form of a Postscript file and as an HTML version.

3. Applications

GraphEnt can recognize and automatically perform several of the most common types of crystallographic syntheses, as discussed in §2.4. Additionally, any type of synthesis that can be reduced to one of these can also be performed, but the reduction step is the responsibility of the user.

As an example of the application of the program, we present results from an anomalous Patterson function calculation using data collected from a horse heart myoglobin crystal. The data were collected with Cu $K\alpha$ radiation. The anomalous signal comes from the iron atom of heme (with $\Delta f''_{\mathrm{Fe,Cu}\,K\alpha} = 3.2\,\mathrm{e}^-$). Table 1 presents statistical information about this data set. To make the example more realistic, we used only data between 20 and 3 Å resolution, and we simulated the presence of outliers in the data by multiplying the amplitude (ΔF_{ano}) and standard uncertainty [$\sigma(\Delta F_{\mathrm{ano}})$] of three randomly chosen strong reflections by a factor of 3.0.

A comparison of the Harker sections (v = 1/2) from the conventional and *GraphEnt* maps (two uppermost panels in Fig. 2) is rather striking: the presence of outliers in the data has completely wiped out the signal from the conventional map, leaving behind a checkerboard appearance, which is all too familiar to macromolecular crystallographers. In sharp contrast, the *GraphEnt* map resembles more a map calculated with hypothetical error-free data than an anomalous Patterson function calculated with real data.

To make the comparison with the *GraphEnt* map more meaningful, we also present (lowest panel of Fig. 2) the same Harker section from a conventional map calculated after rejection of the three outliers.[2] Although this time the Fe–Fe peak is (reassuringly) the strongest peak in the conventional map as well, the *GraphEnt* map (which was calculated with the 'outliers' included in the data) is still by far superior. The difference in the appearance of the two syntheses is not the result of a uniform reduction of the contrast of the *GraphEnt* map: as the two insets in Fig. 2 show, the peak of the *GraphEnt* synthesis stands at approximately 22σ above the mean density of the map, whereas the same peak from the conventional synthesis is only 10σ above the mean.

In summary, the comparison of the conventional and *GraphEnt* maps illustrates all the advantages of the maximum-entropy formalism that are usually cited in the literature: (i) the maximum-entropy map, by being the most uniform map consistent with the observations, only shows detail for which there is evidence in the data, (ii) the effects arising from the presence of outliers in the data are greatly reduced, (iii) the noise level and side lobes (due to series termination errors) are greatly reduced, (iv) the map is everywhere positive and as smooth as the data allow, and (v) the maximum resolution consistent with the data is achieved.

[2] The word 'outlier' is used here catachrestically: as long as the standard uncertainties are correctly estimated, there is nothing wrong with the measurements of these reflections. The common macromolecular practice to exclude large differences from the calculation of Patterson functions arises from the inability of the conventional synthesis to deal correctly with incomplete and noisy data.

4. Program availability

The source code of the program, together with its documentation and some example scripts, is distributed free of charge to both academic and non-academic users, and is immediately available for download *via* http://origin.imbb.forth.gr:8888/~glykos/. The distribution also contains stand-alone executable images suitable for the Silicon Graphics and DEC Alpha OSF workstation architectures.

References

Bricogne, G. (1984). *Acta Cryst.* A**40**, 410–445.
Bricogne, G. (1997). *Methods Enzymol.* **276**, 361–423.
Collaborative Computational Project, Number 4 (1994). *Acta Cryst.* D**50**, 760–763.
Collins, D. M. (1982). *Nature (London)*, **298**, 49–51.
Frigo, M. & Johnson, S. G. (1998). *Proc. ICASSP* **3**, 1381–1384. [Also *via* http://www.fftw.org/fftw-paper-icassp.pdf.]
Gilmore, C. J. (1996). *Acta Cryst.* A**52**, 561–589.
Gull, S. F. & Daniell, G. J. (1978). *Nature (London)*, **272**, 686–690.
Howell, P. L. & Smith, G. D. (1992). *J. Appl. Cryst.* **25**, 81–86.
Jaynes, E. T. (1957). *Phys. Rev.* **106**, 620–630.
Smith, C. R. & Grandy, W. T. Jr (1985). Editors. *Maximum Entropy and Bayesian Methods in Inverse Problems*. Dordrecht: Reidel.
Skilling, J. (1984). *Nature (London)*, **309**, 748–749.
Uffink, J. (1995). *Stud. Hist. Philos. Sci.* **26B**, 223–261. [Also *via* http://www.phys.uu.nl/~wwwgrnsl/jos/mepabst/mepabst.html.]