Acta Crystallographica Section A
Foundations of Crystallography
ISSN 0108-7673

Received 1 September 2005
Accepted 10 November 2005

© 2006 International Union of Crystallography
Printed in Great Britain – all rights reserved

Informational entropy of Fourier maps

Amador Menéndez-Velázquez[a,b]* and Santiago García-Granda[c]

[a]Instituto de Ciencia de Materiales de Madrid, CSIC, Cantoblanco Ctra de Colmenar Km 15, 28049 Madrid, Spain, [b]SpLine, European Synchrotron Radiation Facility, 6 rue Jules Horowit, BP 220, F-38043 Grenoble CEDEX, France, and [c]Departamento de Química Física y Analítica, Facultad de Química, Universidad de Oviedo, 33006 Oviedo, Spain. Correspondence e-mail: amador@icmm.csic.es

An analysis on what is known as the interpretation of Fourier maps has been done from the information theory point of view: determining the nature of the peaks in the map (in order to assign them a suitable scattering factor) and allocating bonds between some of the possible peak pairs. Before interpreting the map, a quantitatively measurable entropy (uncertainty, unknowingness) relating to the molecular structure is known. After the interpretation, this entropy becomes amount of information. This analysis allows us, for the first time in crystallography, to quantify these parameters and analyse the contributions of the different information sources.

1. Introduction

The information theory (Ash, 1967; Pierce, 1980; Gray, 1990; Kahre, 2002; Mackay, 2003; Goldman, 2005) takes care of the problems arising from the conservation, transformation and transmission of information. To deal with these problems mathematically, it is necessary to define, first of all, a *measure for information*.

The information theory is, in fact, a theory of the measure of information, based on the assumption that information is measurable. The information theory has broad applications in cybernetics, telecommunications engineering, information technology, thermodynamics, astronomy, amongst other subject areas.

Attempts to use widely the ideas of information theory in different branches of science are linked to the fact that it is primarily a mathematical theory. Its main concepts (entropy and amount of information) are only determined through probability events, to which the most wide-ranging physical meaning can be attributed.

Among the previous applications of information theory to crystallography, we should mention the work by Diamond (1963), who introduced a measure (in *bits*) of the information contained in the inequality of Karle–Hauptman determinants. Hosoya & Tokonami (1967) considered the estimation of the conformational entropy of an essentially one-dimensional real structure and the removal of structural uncertainty during crystal structure determination by the information contained in the reflection intensities and in the Patterson peaks. de Rango *et al.* (1974) stressed the relation between the efficiency of the probability laws for phase determination and information theory. Gassmann (1977) discussed the problem of coding structural information in crystallography. Piro (1983) derived the information content of some invariants and quantified the gain in information due to *a priori* structural knowledge of the signs of some triplet products.

We shall use the information theory in the crystallography field in order to measure the amount of information that a Fourier map contains.

2. The amount of information

2.1. Concept

In informal language, *information* is a synonym of surprise or knowledge, and is measured by the degree of surprise that it produces on whoever receives it or by the value attributed to it by whoever has it. Such a concept is also used in information theory, although the basic idea behind the classic information theory must rather be acknowledged as the *amount of information* concept.

It is not so obvious that dealing with the amount of information contained in a given message or wanting to measure it with a simple number makes sense. If we want to introduce such a measure, we have to extract the information's shape as well as its content. We should act just like a telegraph office employee who, in order to calculate the sum that must be paid, only takes into account the number of words within the telegraph.

It is appropriate to measure the amount of information contained in a given message by the number of signs that must be used to express its content in the briefest possible way. Any kind of sign system can be used, but the information that has to be measured must then be translated into this system. For technologically imperative reasons, the binary system has been imposed.

In turn, the amount of information within a given message is measured by the number of signs that are needed to express it

by zeros and ones. In this sense, messages that are different in shape and content with respect to the information they contain can thus be compared.

When a number can adopt a value zero or one, the information that indicates that it takes one of these values will be taken as the information unit. The information unit is called a *bit*, an abbreviation of *binary digit*. We will interpret the meaning of one bit as the simplest way to give information: choosing one from two possible alternatives.

2.2. Hartley formulation

If we have a set E of N elements, the formula

$$I(E_N) = \log_2 N, \tag{1}$$

where $I(E_N)$ stands for the information that is needed to define the elements of a set E of N elements, is known as the *Hartley formula* (Hartley, 1928).

The measure of information proposed by Hartley had a tremendous historic importance, as it represented the first equation readily available to measure the amount of information. However, it is necessary to warn that this formula is only valid for a model of information source that attributes equal probabilities to all of its states.

2.3. Shannon formulation

The degree of indetermination when testing a given source outcome depends not only on the number of states but also on the test probabilities. The Shannon formula (Shannon, 1948*a*,*b*, 1949) is more general and allows source events with equally likely states as well as sources with non-equally likely states to be treated, as it incorporates the probabilities directly into the formula. For this reason, this formula is used a lot more in the information theory field than the Hartley formula.

Consider $E_1, E_2, \ldots, E_n$ two by two independent sets, this is

$$E_i \cap E_j = \emptyset, \quad i \neq j, \tag{2}$$

in which

$$E = E_1 \cup E_2 \cup \ldots \cup E_k \cup \ldots \cup E_n, \tag{3}$$

where N_k is the number of elements of E_k. The number of elements of E is given by $N = \sum_{k=1}^{n} N_k$, $p_k = N_k/N$ ($k = 1, 2, \ldots, n$) being the probability that an element chosen at random belongs to set k. In turn, the *Shannon formula* establishes that

$$I = -\sum_{k=1}^{n} p_k \log_2 p_k. \tag{4}$$

Shannon showed that the function I (i) is positive, (ii) increases with increasing uncertainty and (iii) is additive for independent sources of uncertainty.

If, in particular, $n = N$ and $p_1 = p_2 = \ldots = p_N = 1/N$ (equal likely outcomes), the Shannon formula

$$I = -\sum_{k=1}^{N} p_k \log_2 p_k = -\sum_{k=1}^{N} \frac{1}{N} \log_2 \frac{1}{N} = \log_2 N \tag{5}$$

reduces to the Hartley formula.

In this article, we will use either the Hartley formula or the Shannon formula, depending on which one is more suitable to the particular case under study.

3. Amount of information and entropy

3.1. The concept of informational entropy

If we carry out an experiment, the possible outcomes of which are described by the given scheme A, then in doing so we obtain some information (*i.e.* we find out which of the events A_k actually occurs) and the uncertainty of the scheme is completely eliminated. Thus, we can say that the information given to us by carrying out some experiment consists in removing the uncertainty that existed before the experiment. The larger this uncertainty, the larger we consider to be the amount of information obtained by removing it. Shannon referred to the uncertainty of a process as its *entropy*, $S(A)$.

It is natural to express the entropy removed, $S(A)$, by increasing the function of the quantity of information, $I(A)$. The choice of this function means the choice of some unit for the entropy and is therefore fundamentally a matter of indifference. However, it is very convenient to take this entropy proportional to the quantity of information. Of course, the constant of proportionality can be taken as unity, since this choice corresponds merely to a choice of units. Thus, in all that follows, we can consider the amount of information given by the realization of a finite scheme to be equal to the entropy of the scheme. This stipulation makes the concept of entropy especially significant for information theory. Therefore, the amount of information given by the interpretation of the Fourier map will equal the amount of uncertainty that existed before the interpretation.

Because entropy and the total amount of information definitions inform us on the same reality, it is common to hear the term *informational entropy* when referring to the degree of uncertainty of a certain experiment.

3.2. Relationship between entropy and a probability distribution

In this section, we look thoroughly at the relationship between entropy and a probability distribution.

In the probability theory, a complete system of events $\{A_1, A_2, \ldots, A_n\}$ is a set of events in such a way that one and only one of the probability events can occur in a given test or experiment. With this system of events, we associate the set of probabilities $\{p_1, p_2, \ldots, p_i, \ldots, p_n\}$, generating in this way a finite diagram of probabilities. Every finite scheme describes a state of uncertainty. We have an experiment, the outcome of which can be one of the events $\{A_1, A_2, A_3, \ldots, A_n\}$ and we know only the probabilities of these possible outcomes. It seems obvious that the amount of uncertainty is different in different schemes. Thus, in the two simple alternatives

$$\begin{bmatrix} A_1 & A_2 \\ 0.5 & 0.5 \end{bmatrix}, \quad \begin{bmatrix} A_1 & A_2 \\ 0.99 & 0.01 \end{bmatrix}, \tag{6}$$

the first obviously represents more uncertainty than the second; in the second case, the result of the experiment is 'almost surely' A_1, while in the first case we naturally refrain from making any prediction. The scheme

$$\begin{bmatrix} A_1 & A_2 \\ 0.3 & 0.7 \end{bmatrix} \tag{7}$$

represents an amount of uncertainty intermediate between the preceding two.

Shannon's definition of the total amount of information I or informational entropy S associated with a probability distribution $P = \{p_1, p_2, \ldots, p_i, \ldots, p_n\}$,

$$I(P) = S(P) = -\sum_{k=1}^{n} p_k \log_2 p_k, \tag{8}$$

is in agreement with the previous probability interpretation, as it assigns more entropy to the greatest uncertainty alternative. Thus, for example, if we apply equation (8) to the three previous probability diagrams, we obtain as informational entropy values 1, 0.081 and 0.88, respectively, which confirms that an increase in uncertainty results in an increase of informational entropy.

According to the previous interpretation, the maximum-entropy value (maximum uncertainty, maximum unknowingness) will be attained when the possible results of the random experiment are equally likely and will be represented by

$$S_{\max} = \log_2 N. \tag{9}$$

Entropy's minimum value will be null,

$$S_{\min} = 0, \tag{10}$$

this means that there is no indetermination upon the result of a given experiment.

The applicability of the information theory to different systems or structures is only conditioned by the possibility of building a finite diagram of probabilities. If this is possible, equation (8) will then take us directly to the amount of information or informational entropy.

4. Informational entropy of Fourier maps

4.1. Approaching the problem

Since 1948, the year in which Shannon analysed the statistical consequences of entropy and applied it to the communications field, numerous physicists have attempted to find applications for informational entropy. For example, Jaynes (1957) used the ideas of information theory to build the fantastic world of statistical mechanics. In this section, we use the information theory in order to calculate the information contained in a Fourier map.

A Fourier map may be regarded as a particular set of peaks pertaining to a set of many possible crystal structures. Therefore, an entropy (uncertainty, ignorance) might be defined for such a map. In order to solve the structure, this uncertainty must be removed. This information should match the entropy.

Let us consider a Fourier map of N peaks, $A_1, A_2, \ldots, A_N$, and n scattering-factor types [corresponding to a (sub)molecule of N atoms and n different atomic species]. By setting this hypothesis, we are assuming that there are no spurious peaks, so it does not involve any restriction. If we want to include the spurious peaks, we could incorporate them into the diagram that we develop but consider them as an additional type of atom that does not chemically bond to any of the other atoms. Interpreting a Fourier map means, on the one hand, identifying each peak in the map with a certain atomic species (which is essentially equal to assigning to each peak a certain scattering factor) and, on the other hand, establishing or not a chemical bond between each of the possible pairs of peaks. We have, in turn, two different types of uncertainty in the Fourier map, which will give rise to two entropy types. We will call the first one *informational entropy due to scattering factor*, $S(S)$, and the second one *informational entropy due to connectivity*, $S(C)$.

4.2. Informational entropy due to connectivity

Let us consider now the informational entropy due to connectivity. This entropy is caused by the uncertainty present due to the existence or not of chemical bonds between each of the possible pairs of atoms.

Suppose that for any two atoms (peaks in the Fourier map) we have to decide which of the following statements is correct:

(*a*) In the molecule M, there is a chemical bond between the atoms A_i and A_j.

(*b*) In the molecule M, the atoms A_i and A_j are not chemically bonded.

Every unordered pair of atoms, (A_i, A_j), is a discrete information source. The number of theoretically possible bonds in a (sub)molecule is $N(N-1)/2$, since $a_{ii} = 0$ and $a_{ij} = a_{ji}$, a_{ij} being the ijth element of the connectivity matrix. The probability that a randomly chosen pair of atoms are chemically bonded is

$$p_b = \tfrac{1}{2} \tag{11}$$

and the informational entropy in connectivity for such a (sub)molecule, $S(C)$ is

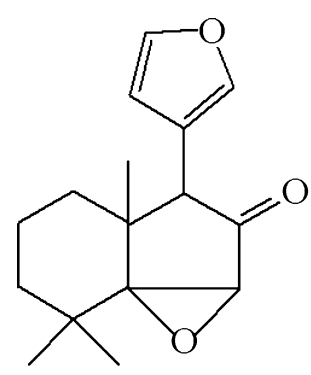

Figure 1
Molecular structure of compound X.

$$S(C) = -\sum_{i=1}^{NE} (p_b \log_2 p_b + \bar{p}_b \log_2 \bar{p}_b) = \frac{N(N-1)}{2}, \quad (12)$$

where

$$\bar{p}_b = 1 - p_b = \tfrac{1}{2} \quad (13)$$

and being $NE = N(N-1)/2$, the number of edges (possible bonds) of the complete graph with N vertices, K_N.

It is interesting to note how this derived entropy shows a maximum-entropy value in the Shannon formula. As previously argued, this is because the maximum-entropy value is obtained when the possible states of an information source are equally likely, just as we have assumed in our formulation by assigning the same probability to the existence or non-existence of a chemical bond.

As we are considering equal probabilities for the existence or non-existence of a chemical bond, we can calculate informational entropy according to the Hartley formula. If we consider a pair of atoms as an information source, there are only two possibilities related to the connectivity (bond or no bond). In turn,

$$S_b = \log_2 2 = 1, \quad (14)$$

where S_b is the entropy due to a pair of atoms or potential bond. Owing to the additions of independent information sources, the entropy due to connectivity of a Fourier map will be given by

$$S(C) = \sum_{i=1}^{NE} 1 = NE = \frac{N(N-1)}{2}, \quad (15)$$

obtaining the same result as previously through the Shannon formula.

We can also calculate entropy according to the Hartley formula but following a different approach to the one we have just developed. Now, we shall not consider a pair of atoms or peaks in the map as an information source but all of the peaks from the map. Given a set of peaks in the map (corresponding to atomic nuclei), we can establish different connectivity relationships amongst them. Each network of bonds will give rise to a certain molecular structure. Hartley's different possibilities will now be the possible molecular structures that can be generated from a set of N peaks. The number of possible molecular structures for a set of N peaks can be calculated[1] (Menéndez-Velázquez, 1999; Menéndez-Velázquez & García-Granda, 2006). In the case of non-labelled peaks, this number will be given by $2^{NE} = 2^{N(N-1)/2}$. In turn, if there are $2^{N(N-1)/2}$ possibilities of equal probability, the informational entropy of the map associated with the connectivity will be given by

$$S(C) = \log_2 2^{N(N-1)/2} = N(N-1)/2. \quad (16)$$

We can therefore see how, by using the different methods, we can reach the same informational entropy value. In all of these cases, the informational entropy due to connectivity corresponds to a Fourier map with labelled peaks.

If we considered the case of non-labelled peaks, the informational entropy of a Fourier map would decrease, as there would be pairs of atoms or bonds potentially equivalent by symmetry and, as a consequence, the number of existing unknowns or uncertainty in the map relating to a bond would be reduced. If we have, for example, two pairs of atoms equivalent by topological symmetry, it would be enough to store the information of one of the pairs, avoiding in this way doubling up the information. In order to consider the case of non-labelled peaks, the most practical thing to do is to use the Hartley formula, as then we will only need to know the number of potential molecular structures non-equivalent by symmetry in a Fourier map of N peaks[2] (Menéndez-Velázquez, 1999; Menéndez-Velázquez & García-Granda, 2006). In Tables 1 and 2, we show the informational entropy due to connectivity, $S(C)$, of a Fourier map with different numbers of peak values, and considering the case of labelled peaks as well as that of non-labelled peaks. Here we can clearly see that, when considering non-labelled peaks, entropy (uncertainty) decreases, as then pairs of atoms that are topologically equivalent start to appear, thus reducing the number of different pairs of atoms upon which there is an uncertainty relating to the existence or non-existence of a chemical bond.

4.3. Informational entropy due to the scattering factor

Once having considered the informational entropy due to connectivity, let us now have a look at the informational entropy due to the scattering factor. Let us consider first the case of labelled peaks.

Let us suppose, as an example, that the molecular structure to elucidate is that of compound X shown in Fig. 1.

Before solving the structural make-up, the previous analysis methods tell us that the empirical formula of the compound to be elucidated is $C_{16}H_{20}O_3$. Let us now suppose that, in the process of finding the structural composition, we generate a Fourier map with peaks corresponding to all the atoms, except H atoms, in the molecular structure to be elucidated. Let us also suppose that there are no spurious peaks present. With these preliminary hypotheses, the Fourier map will consist of 19 peaks.

The sources of information associated with the scattering factor are not independent information sources, as we will prove here. In the example that we are considering of compound X, there are 19 atoms, 3 of which are O and the rest

[1] The combinatorial algebraic, through the Polya theorem and its generating functions, provides adequate tools to carry out an account of molecular graphs with N nodes and a different number B of bonds, B taking the values $B = 0, 1, 2, \ldots, N(N-1)/2$. Adding the number of (N, B) graphs – graphs with N nodes and B bonds – for the different values of B, we obtain the total number of possible molecular graphs that can be generated from N nodes through the assignation of bonds. This last number corresponds to the cardinal of E set, $|E|$, in the Hartley formulation.

[2] If peaks are not labelled, it is necessary to take into account the topological symmetry of the graphs to avoid repeatedly counting the graphs that are equivalent for their symmetry. This is possible through the introduction of automorphism groups and the so-called cycle index that reintroduce in a way the symmetry in the Polya theorem.

Table 1
Informational entropy due to connectivity, $S(C)$, for various Fourier maps with different numbers of labelled peaks, $|E|$ being the number of elements in the set E (Hartley formulation).

No. of labelled peaks	$\|E\|$ number of possible labelled molecular graphs	$S(C)$ (labelled peaks)
1	1	0
2	2	1
3	8	3
4	64	6
5	1 024	10
6	32 768	15
7	2 097 152	21
8	268 435 456	28
9	68 719 476 736	36
10	35 184 372 088 832	45
11	36 028 797 018 963 968	55
12	73 786 976 294 838 206 464	66
13	302 231 454 903 657 293 676 544	78
14	2 475 880 078 570 760 549 798 248 448	91
15	40 564 819 207 303 340 847 894 502 572 032	105
16	1 329 227 995 784 915 872 903 807 060 280 344 576	120
17	87 112 285 931 760 246 646 623 899 502 532 662 132 736	136
18	11 417 981 541 647 679 048 466 287 755 595 961 091 061 972 992	153
19	2 993 155 353 253 689 176 481 146 537 402 947 624 255 349 848 014 848	171
20	1 569 275 433 846 670 190 958 947 355 801 916 604 025 588 861 116 008 628 224	190

Table 2
Informational entropy due to connectivity, $S(C)$, for various Fourier maps with different numbers of non-labelled peaks, $|E|$ being the number of elements in the set E (Hartley formulation).

No. of non-labelled peaks	$\|E\|$ number of possible non-labelled molecular graphs	$S(C)$ (non-labelled peaks)
1	1	0
2	2	1
3	4	2
4	11	3.4594
5	34	5.0875
6	156	7.2854
7	1 044	10.0279
8	12 346	13.5918
9	274 668	18.0673
10	12 005 168	23.5171
11	1 018 997 864	29.9245
12	165 091 172 592	37.2645
13	50 502 031 367 952	45.5214
14	29 054 155 657 235 488	54.6896
15	31 426 485 969 804 308 768	67.7686
16	64 001 015 704 527 557 894 928	75.7605
17	245 935 864 153 532 932 683 719 776	87.6684
18	1 787 577 725 145 611 700 547 878 190 848	100.4960
19	24 637 809 253 125 004 524 383 007 491 432 768	114.2464
20	645 490 122 795 799 841 856 164 638 490 742 749 440	128.9237

are C atoms. Initially, an atom chosen at random has a probability of 3/19 of being an O atom and a probability of 16/19 of being a C atom. If we now proceed onto choosing a second atom (allocating its corresponding scattering factor), we must then modify the previous probability diagram. Besides, the new probability diagram that we must generate depends on the atomic nature of the first atom. If the first atom was a C atom, the probability diagram for the second atom, randomly chosen, is

$$\begin{pmatrix} \mathrm{C} & \mathrm{O} \\ 15/18 & 3/18 \end{pmatrix}, \qquad (17)$$

whereas if the first atom was an O atom, the corresponding probability diagram for the second atom is

$$\begin{pmatrix} \mathrm{C} & \mathrm{O} \\ 16/18 & 2/18 \end{pmatrix}. \qquad (18)$$

We are, in turn, facing a conditional probability, given that the probability of an event happening depends on the verification or not of other events. If we wanted to calculate the informational entropy according to this method, we would have to consider all these possibilities, which implies a long and tedious process. Let us try then to calculate the informational entropy due to the scattering factor through another method, specifically through the Hartley formulation.

In order to apply the Hartley formulation, we must have an information source with equally likely chances. If we consider each peak of the map separately, we are facing non-equally likely information sources, as we have seen before, and then the Hartley formulation is not valid.

In turn, we must consider as an information source the whole group of 19 labelled peaks of the Fourier map and then proceed to count the different possible nuclear configurations resulting from a particular allocation of a given atomic species (carbon or oxygen) to each of the 19 peaks in the map. Each of the possible nuclear configurations created represents a possible state and where all the states are equally likely, so that the Hartley formulation can then be applied.

The problem that arises is equivalent to the following. We have 19 free boxes and we have to place a single atom, either a C or an O atom, in each of the boxes, bearing in mind that we have 16 C atoms and 3 O atoms. In how many ways can we place the 3 O atoms and the 16 C atoms in the 19 boxes, knowing that in each box we can only introduce one atom? This is a simple combinational problem. The answer is given by the number of permutations with 19 elements, out of which 16 are of one type and 3 are of another type, $PR_{19}^{16,3} = 969$. According to the Hartley formulation, the informational entropy due to the scattering factor, $S(S)$, is given by

Table 3
Informational entropy due to the scattering factor, $S(S)$, for the case of 19 atoms and different empirical formulas (excluding the H atoms), $|E|$ being the number of elements in the set E (Hartley formulation).

Empirical formula	$\|E\|$	$S(S)$	Empirical formula	$\|E\|$	$S(S)$
C_{19}	1	0	$C_{17}OS$	342	8.4179
$C_{18}O$	19	4.2479	$C_{16}O_2S$	2 907	11.5053
$C_{17}O_2$	171	7.4179	$C_{15}O_3S$	15 504	13.9204
$C_{16}O_3$	969	9.9204	$C_{15}O_2S_2$	23 256	14.5053
$C_{15}O_4$	3 876	11.9204	$C_{14}O_4S$	58 140	15.8272
$C_{14}O_5$	11 628	13.5053	$C_{14}O_3S_2$	116 280	16.8272
$C_{13}O_6$	27 132	14.7277	$C_{13}O_5S$	162 792	17.3127
$C_{12}O_7$	50 388	15.6208	$C_{13}O_4S_2$	406 980	18.6346
$C_{11}O_8$	75 582	16.2058	$C_{13}O_3S_3$	542 640	19.0496
$C_{10}O_9$	92 738	16.4953	$C_{12}O_6S$	352 716	18.4281
C_9O_{10}	92 738	16.4953	$C_{12}O_5S_2$	1 058 148	20.0131
C_8O_{11}	75 582	16.2058	$C_{12}O_4S_3$	1 763 580	20.7501
C_7O_{12}	50 388	15.6208	$C_{11}O_7S$	604 656	19.2058
C_6O_{13}	27 132	14.7277	$C_{11}O_6S_2$	2 116 296	21.0131
C_5O_{14}	11 628	13.5053	$C_{11}O_5S_3$	4 232 592	22.0131
C_4O_{15}	3 876	11.9204	$C_{11}O_4S_4$	5 290 740	22.3350
C_3O_{16}	969	9.9204	$C_{10}O_8S$	831 402	19.6652
C_2O_{17}	171	7.4179	$C_{10}O_7S_2$	3 325 608	21.6652
CO_{18}	19	4.2479	$C_{10}O_6S_3$	7 759 752	22.8876
O_{19}	1	0	$C_{10}O_5S_4$	11 639 628	23.4725

$$S(S) = \log_2 N = \log_2 969 = 9.9204. \quad (19)$$

The case that we have just developed corresponds to the case of labelled peaks. If the peaks were non-labelled, some of the possible vertex-weighted molecular graphs, resulting after the allocation of atomic species on specific molecular graphs (network of bonds), would be equivalent to each other by topological symmetry and could thus be eliminated and, as a consequence, the number of elements of set E to which Hartley makes reference, and the resulting informational entropy would take lower values. Despite this, values of informational entropy due to the scattering factor are much lower that the corresponding values of informational entropy due to connectivity. This is one of the really interesting conclusions that we were looking for.

In Table 3, we show the informational entropy due to the scattering factor of a Fourier map of 19 peaks, for different empirical formulas. In order to simplify the problem, we have only considered the case of labelled peaks, which represents the upper limit of informational entropy (unknowingness) associated with the scattering factor.

4.4. Total informational entropy of Fourier maps

The total informational entropy of the Fourier map, $S(F)$, will be the sum of the informational entropies due to connectivity and due to the scattering factor,

$$S(F) = S(C) + S(S). \quad (20)$$

If we consider the case of compound X, whose empirical formula excluding H atoms is $C_{16}O_3$, the informational entropy of the corresponding associated Fourier map will be given by (labelled peaks scenario)

$$S(F) = S(C) + S(S) = 171 + 9.9204 = 180.9204 \quad (21)$$

or by

$$S(F) = S(C) + S(S) \leq 114.2464 + 9.9204 = 124.1668 \quad (22)$$

in the case of non-labelled peaks.

As we can see, the informational entropy takes very high values. Besides, the entropy due to connectivity exceeds the entropy due to the scattering factor. This difference and these high values shoot up as we increase the number of peaks in the map. For example, if we consider the case of a compound of empirical formula $C_{90}O_{10}$, the informational entropy of the associated Fourier map will be given by (labelled peaks scenario)

$$S(F) = S(C) + S(S) = 4950 + 43.9767 = 4993.9767. \quad (23)$$

This entropy (unknowingness, uncertainty) due to connectivity is unnecessary and may be removed if we use a good method for interpreting Fourier maps, such as topological analysis, which can be performed in any Fourier map as far as classical peak-picking works, if a robust procedure to extract all the critical points is available (Menéndez-Velázquez, 1999; Menéndez-Velázquez & García-Granda, 2003).

5. Concluding remarks

The interpretation of a Fourier map from the information theory point of view has been analysed by making use of both Hartley and Shannon formulations. This analysis has allowed us to conclude that there is a large informational entropy (synonymous to uncertainty, unknowingness) in a Fourier map, as a result of the different possible allocations of scattering factors to the peaks in the map and due to the different possible allocations of bonds between pairs of atoms. From this analysis, we also found that entropy increases remarkably as the number of map peaks increases and that the uncertainty related to the chemical bond is much bigger than the uncertainty due to the atomic nature.

SGG acknowledges MCyT (BQU2003-05093) for financial support.

References

Ash, R. (1967). *Information Theory*. New York: John Wiley and Sons.
Diamond, R. (1963). *Acta Cryst.* **16**, 627–639.
Gassmann, J. (1977). *Acta Cryst.* A**33**, 474–479.
Goldman, S. (2005). *Information Theory*. New York: Dover Phoenix Editions.
Gray, R. M. (1990). *Entropy and Information Theory*. New York: Springer-Verlag.
Hartley, R. V. L. (1928). *Bell System Tech. J.* **7**, 535.
Hosoya, S. & Tokonami, M. (1967). *Acta Cryst.* **23**, 18–25.
Jaynes, E. T. (1957). *Phys. Rev.* **106**, 620–630.
Kahre, J. (2002). *The Mathematical Theory of Information*. Boston: Kluwer Academic Publishers.
Mackay, D. J. C. (2003). *Information Theory, Inference and Learning Algorithms*. Cambridge University Press.
Menéndez-Velázquez, A. (1999). PhD thesis, University of Oviedo, Spain.

Menéndez-Velázquez, A. & García-Granda, S. (2003). *J. Appl. Cryst.* **36**, 193–205.
Menéndez-Velázquez, A. & García-Granda, S. (2006). In preparation.
Pierce, J. R. (1980). *An Introduction to Information Theory. Symbols, Signals and Noise.* New York: Dover Publications.
Piro, O. E. (1983). *Acta Cryst.* A**39**, 61–68.
Rango, C. de, Tsoucaris, G. & Zelwer, C. (1974). *Acta Cryst.* A**30**, 342–353.
Shannon, C. E. (1948*a*). *Bell Syst. Tech. J.* **27**, 379.
Shannon, C. E. (1948*b*). *Bell Syst. Tech. J.* **27**, 623.
Shannon, C. E. (1949). *Bell Syst. Tech. J.* **28**, 656.