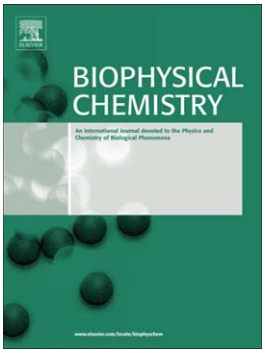
Criteria for confirming sequence periodicity identified by Fourier transform analysis: application to GCR2, a candidate plant GPCR?

Christopher J.R. Illingworth, Kevin E. Parkes, Christopher R. Snell, Philip M. Mullineaux, Christopher A. Reynolds



Criteria for confirming sequence periodicity identified by Fourier transform analysis: application to GCR2, a candidate plant GPCR?

Christopher J.R. Illingworth, Kevin E. Parkes, Christopher R. Snell, Philip M. Mullineaux, Christopher A. Reynolds

PII: S0301-4622(07)00277-3
DOI: doi: 10.1016/j.bpc.2007.11.004
Reference: BIOCHE 5047

To appear in: *Biophysical Chemistry*

Received date: 26 September 2007
Revised date: 15 November 2007
Accepted date: 15 November 2007

Please cite this article as: Christopher J.R. Illingworth, Kevin E. Parkes, Christopher R. Snell, Philip M. Mullineaux, Christopher A. Reynolds, Criteria for confirming sequence periodicity identified by Fourier transform analysis: application to GCR2, a candidate plant GPCR?, *Biophysical Chemistry* (2007), doi: 10.1016/j.bpc.2007.11.004

Criteria for confirming sequence periodicity identified by Fourier transform analysis: application to GCR2, a candidate plant GPCR?

Christopher J. R. Illingworth[1], Kevin E. Parkes[2], Christopher R. Snell[2], Philip M. Mullineaux[1] and Christopher A. Reynolds*[,1]

[1]Department of Biological Sciences, University of Essex, Wivenhoe Park, Colchester, CO4 3SQ, United Kingdom. [2] *Medivir UK Ltd., Chesterford Research Park, Little Chesterford, Essex CB10 1XL, United Kingdom.*

Abstract

Methods to determine peridiodicity in protein sequences are useful for inferring function. Fourier transformation is one approach but care is required to ensure the periodicity is genuine. Here we have shown that empirically-derived statistical tables can be used as a measure of significance. Genuine protein sequences data rather than randomly generated sequences were used as the statistical backdrop. The method has been applied to G-protein coupled receptor (GPCR) sequences, by Fourier transformation of hydrophobicity values, codon frequencies and the extent of over-representation of codon pairs; the latter being related to translational step times. Genuine periodicity was observed in the hydrophobicity whereas the apparent periodicity (as inferred from previously reported measures) in the translation step times was not validated statistically. GCR2 has recently been proposed as the plant GPCR receptor for the hormone abscisic acid. It has homology to the Lanthionine synthetase C-like family of proteins, an observation confirmed by fold recognition. Application of the Fourier transform algorithm to the GCR2 family revealed strongly predicted seven fold periodicity in hydrophobicity, suggesting why GCR2 has been reported to be a GPCR, despite negative indications in most transmembrane prediction algorithms. The underlying multiple sequence alignment, also required for the Fourier transform analysis of periodicity, indicated that the hydrophobic regions around the 7 GXXG motifs commence near the C-terminal end of each of the 7 inner helices of the α-toroid and continue to the N-terminal region of the helix. The results clearly explain why GCR2 has been understandably but erroneously predicted to be a GPCR.

Keywords: Fourier transform, periodicity, codon pairs, codons, G-protein coupled receptors, hydrophobicity.

Introduction

Some degree of symmetry and periodicity is occasionally observed in protein sequence and structure and this is often related to function. Here we present criteria to ensure that periodicity inferred from Fourier transform approaches is not over reported, and apply these methods to G-protein coupled receptors (GPCRs). Discrete Fourier transformation is one of many methods that can be used to infer structure and function from the physical properties associated with a protein sequence. The importance of such methods arises from the need to analyze the ever growing wealth of protein sequence data arising through genome projects. Fourier analysis is particularly well-suited to looking for patterns within the amino-acid sequences. In one of the earlier predictions of symmetry, Zimmerman used Fourier transforms and an autocorrelation function to search for periodicities in residue properties such as volume and interchangeableness, and inferred a 5-residue repeating pattern in the polarity of the residues in the bakers yeast cytochrome c sequence [1] , which may be related to stretches of small amino acids in the alpha helices of the protein structure. MacLachlan and Stewart used Fourier transforms to find a 14-fold periodicity in α-tropomysin, and demonstrated the statistical significance of this result through a mathematical analysis of the Fourier transform method [2] ; the 14-fold periodicity was later confirmed by X-ray crystallography [3,4]. Statistically significant periodicities have also been found through the application of Fourier transform methods to DNA sequences [5-8]. At a more local level of protein structure, Fourier transforms have been successfully applied in conjunction with hydrophobicity scales to reveal amphiphilic secondary structures in protein sequences [9-14].

More recently, weaknesses in the Fourier transform method have been identified, for example in the potential loss of periodicity when a protein is converted into a numerical sequence [15,16], and this has led to the development of other methods of determining periodicity in protein sequences [17,18]. However, the Fourier transform remains a valid method for searching for periodicity in a particular property of a sequence, especially if the property is not necessarily related to the individual amino acids in a simple 1-to-1 fashion (see below).

Another approach to searching for periodicity in amino acid sequences involved the application of Fourier transform methods to sets of proteins, and processing the transformed sequences of individual proteins to generate a combined signal measuring periodicities which are common to the set. Periodicities predicted by this method have been related to both structural factors [19]

and protein function [20]. In these applications of Fourier transforms across sets of proteins, particularly large peaks in the combined signal were taken as indicative of common periodicities, however, there was no attempt to quantify the certainty with which this periodicity can be inferred. Here, Monte Carlo envelopes are used to give, for the first time, indications of the statistical significance of these methods. In an initial test, Fourier transform methods are applied to random protein sequences to study previously reported significance levels. The same methods are then used to study hydrophobicity, and the factors believed to govern the speed of codon translation, in sets of G-protein coupled receptor (GPCR) sequences.

The results highlight potential pitfalls of the method, and suggest that previous predictions of periodicity may have been over-interpreted, though they also illustrate cases in which the method can be very useful, for example in uncovering genuine low frequency periodicities. To illustrate the power of the method, we have applied the method to the GCR2 family.

GCR2 [21] and the homologous Lanthionine synthetase C-like proteins [22-25], have been reported to be GPCRs [21,26,27], primarily because they have been identified by transmembrane helix prediction algorithms as having 7 transmembrane helices, but in the case of the LANCL1 protein, motifs such as putative glycosylation sites were also identified [26,27]. There is currently much interest in GCR2 because it has been proposed as the receptor for abscisic acid [21], an important plant hormone. However, the LANCL1 protein was later re-classified by the original authors as a peripheral membrane protein with enzyme activity [24], an observation now justified by the recent X-ray crystal structure of a lantibiotic cyclase (PDB codes 2g02, 2g0d). Recently the status of GCR2 as a GPCR has also been queried [28] and in an attempt to understand the origin of this confusion, we have analysed the GCR2 - Lanthionine synthetase C-like protein family using Fourier transform analysis, using hydrophobicity as the transformed property.

Methods

Discrete Fourier transform

From any one-dimensional sequence of amino acids of length l, a numerical sequence f(k) can be derived, by assigning numerical values, for example hydrophobicity scores, to the amino acids in the sequence. Given such a numerical sequence, the discrete Fourier transform is the sequence F(n), where k is the position along the numerical sequence and n is the frequency, given by

$$F(n) = \sum_{k=1}^{l} f(k) e^{2\pi i k n / l} \quad (1)$$

F(n) is usually complex, and can be separated into a real cosine series, and an imaginary sine series, as follows:

$$F(n) = C(n) + iS(n) = \sum_{k=1}^{l} f(k)\cos(2\pi k n / l) + i\sum_{k=1}^{l} f(k)\sin(2\pi k n / l) \quad (2)$$

Signal-to-noise ratio, S / N

Given a numerical sequence f(w) with m elements, the signal-to-noise ratio is the maximum absolute value of an element of the sequence divided by the mean absolute value of the sequence

$$S / N = \max_w \{|f(w)|\} / \sum_{w=1}^{m} \frac{|f(w)|}{m} \quad (3)$$

The signal-to-noise ratio has a minimal value of 1, in the case of all elements of f(w) having the same absolute value, and a maximal value of m, in the case of all but one elements of f(w) having an absolute value of zero.

Random sequence generation

A number of different sets of sequences were used in the study. In order to study the significance values quoted by de Trad et al. [29], and Cosic , a large number of random sequences were generated. A set of 100 000 proteins was randomly selected from the UniRef50 database [30], which consists of protein sequences clustered such that no two sequences in the database have more than 50% sequence identity. A random residue was chosen in a random sequence, and the protein sequence was read off from that point. If the end of the sequence was reached, a jump was made to another random sequence from the 100 000, starting at the residue int(p^2L), where L was the length of the new sequence, and p was a uniform random variable on the interval [0,1]. Here the p^2 term biases the choice of residue towards the start of the protein, in order to minimise the number of jumps between sequences. Generating random sequences from real protein data, rather than on a residue by residue basis, incorporates into the random sequences more of the autocorrelations found in real protein sequences than would the generation of sequences on a residue-by-residue basis. Thus, results from specific protein sequences can be compared to a backdrop of what would be expected from a protein sequence that was chosen by chance. Residues from these random sequences were converted into their EIIP (Electron Ion Interaction Potential) values [31], nominally ranging from 0 to 0.1263. The EIIP value, sometimes referred to as the PEII value [32], is based on a pseudopotential method [33], and was claimed to

correlate with properties of organic molecules such as carcinogenicity, toxicity, and antibiotic activity [34-37]. Though this claim was the subject of some controversy [38-39], the method is used here for consistency with the previous key studies in this area. The experiment was also carried out using random sequences generated on a residue-by-residue basis, which by definition have no inherent autocorrelation. Results from these latter tests are contained in supporting information.

GPCR sequence generation

A second set of sequences was used to study potential periodicity in codon and codon pair data and in hydrophobicity. The alignment of olfactory proteins from the GPCR database [40] was edited to extract a set of human olfactory proteins with no gaps or insertions, each of 314 residues in length, and these were filtered using the program Jalview [41] to remove redundancy, so that no two sequences had more than 50% sequence similarity. DNA sequences for each of the resulting 12 proteins were taken from the EMBL nucleotide database [42] and converted into χ^2 values describing the frequency of the DNA codons, using information from the codon usage database (http://www.kazusa.or.jp/codon/), and the extent of over-representation of codon pairs, derived from the work of Gutman [43], to give two numerical sequences describing each protein.

χ^2 values for codons that occur less frequently than average, and for codon pairs that were over-represented, were arbitrarily given a positive sign. Other work, not reported here, has suggested that codons and codon pairs which are translated slowly may have a role in protein folding. Rare codons, and over-represented codon pairs [44] that are reportedly translated slowly [45] and which have positive χ^2 values were the focus of this project, so χ^2-values which were below zero were set to equal zero, representing the assumption that the speed of translation is unimportant when the translation happens quickly. Further sets of sequences were used to study periodicity in the hydrophobicity of residues. Twenty six proteins were randomly selected from the multiple sequence alignment of archaeal bacteriorhodopsins in the GPCR database [46], and another thirty proteins randomly selected from the multiple sequence alignment of all rhodopsin vertebrate sequences in the same database. These were converted into numerical hydrophobicity values according to a measure of the hydrophobicity of each residue [47]. Forty six proteins from a GCR2 alignment were also studied with this hydrophobicity method (see below).

RRM method

In order to analyze each of the numerical sequences, the RRM method described by Cosic [48] was applied to the sequences f(k). In-house code was used to generate the real and imaginary

parts of a Fourier transform (see Eq. 2), which in turn were used to generate the real sequence R(n)

$$R(n) = \sqrt{C(n)^2 + S(n)^2} \quad (4)$$

Given a protein length of L, the sequence R(n) takes values for $0 \leq n \leq L/2$, as any pattern found in the protein could not have a wavelength shorter than 2 residues. Where $R(n)_i$ is the sequence corresponding to protein j, repeating this process for all of the proteins in the set and multiplying together generated the cross-spectral function

$$P(n) = \prod_{j=1}^{M} R(n)_j \quad (5)$$

where the product is taken over all of the M transformed sequences from the protein set. This multiplication identifies frequencies n which have high values of R(n) for most values of j – if for a few values of j, R(n) is small, the product will also be small. Hence the multiplication step identifies common frequencies in the data [49]. As a measure of the significance of the resultant signal, the signal-to-noise ratio, S/N, was calculated for P(n) using Equation 3.3. Where the RRM method was applied to codon and codon pair data, and to residue hydrophobicity scores, in-house code was used to calculate the discrete Fourier transforms, rather than the Fast Fourier Transform (FFT) method used by Cosic [50]. This allows for easier interpretation of frequency results.

Significance tests

In order to provide an estimate of significance levels for the EIIP method, sets of between 1 and 30 random sequences were generated, with lengths varying from 100 to 400 residues. For each length and set size, 10 000 sets of proteins were randomly generated. Calculating the S/N ratio for each of these sets, and ordering them, gave statistically derived estimates for significance at the 50%, 95%, and 99% levels.

In order to test the significance of the result from the olfactory protein DNA data, the RRM method was applied to 10 000 sets of 12 random DNA sequences. Random sequences were generated from DNA taken from proteins in the human genome, from the EMBL nucleotide human coding sequence database, [42] edited to remove duplicate sequences, partial sequences, and sequences with bases other than G, C, A, or T. These sequences were filtered to remove any sequences less than 315 amino acids in length, to give a set of 57 825 proteins. The χ^2 information for the codons and for the codon pairs were derived as above. To generate a random

sequence, a random protein was chosen from the set, and the χ^2 scores for the codons and for the codon pairs were read along the DNA sequence, starting from the first codon of the sequence.

In a similar manner, 10 000 sets of random sequences equal in number and length to the sequences in the hydrophobicity test sets were generated and converted into hydrophobicity scores, to give a statistical indication of the significance of the results obtained.

GCR2 transmembrane helix prediction and sequence alignment

The transmembrane regions of GCR2 and related GPCRs were predicted using TMHMM [51-52] and the Kyte-Doolittle method [53]. The results indicating that GCR2 is not a GPCR are shown in supporting information. The alignment of GCR2 with the lantibiotic cyclase (PDB codes 2g02) [54] and the PFAM [54] seeded alignment of the Lanthionine synthetase C-like proteins (pfam code LANC_like/PF05147**)** was generated using a profile alignment with clustalX [55-56]. This homology of GCR2 to the Lanthionine synthetase C-like proteins has been reported elsewhere [25]. We note that all of the key Lanthionine synthetase C-like GXXG motifs [24] are aligned in both GCR2 and 2g02, along with the catalytic residues of the lantibiotic cyclase (PDB code 2g02)[22]. Hydrophobicity values were assigned to each position, as above [47]. The Fugue [57], genTHREADER [58] and Phyre [59] fold recognition servers all identified lantibiotic cyclase as a high scoring hit for GCR2 (see supporting information).

Results

Results of the measurement of S/N values in the random EIIP sequences are shown in Table 1. The level of significance is dependent on both the number of sequences in a set, and on the length of those sequences. The figures obtained contrast with the S/N value of 20 which, following the work of Veljkovic et al [60] has widely been assumed as being significant [61-65] – for sets of 30 proteins, this value would in fact be below average. In order to demonstrate statistical significance at the 95% level, much higher values would often be needed. Thus to infer periodicity in the EIIP values with a 95% certainty for a set of 20 proteins of length 300 amino acids, a single to noise ratio in excess of 98.2 is required.

Applied to the codon data, the RRM method found a spike in the cross-spectral function with a signal-to-noise ratio of 22.7, at a (very high) frequency of 116 (i.e. there are 116 repeating units in the 314 residues, corresponding to a wavelength of about 2.7 residues). Applied to the codon pair data, the RRM method found a spike with signal-to-noise ratio of 27.3, again at a (very high) frequency of 136. (These two frequencies are 0.3694 and 0.4331 in Cosic's measure.) In both cases, the signal-to-noise ratio was above the value of 20, identified by Cosic as being the threshold for significance. However, application of the RRM method to random DNA sequences suggested another picture with regards to significance. For the codon and codon pair data, the median signal-to-noise ratios from 10 000 sets of random DNA sequences were 24.8 and 23.8 respectively. Within the set of results from random protein sequences, the 5% and 1% high values were, respectively, 59.7 and 85.9 for the codon data, and 57.3 and 82.0 for the codon pair data. Thus, the signal-to-noise values obtained for the olfactory protein DNA data do not appear to be significant, and it is likely that the spikes obtained are the product of chance, rather than any interpretable pattern in the sequence data. Applying the RRM method to the set of bacteriorhodopsin sequences gave a spike in the cross-spectral function with S/N ratio of 114 at a frequency of 7. Applying the same method to the 10 000 sets of random proteins of the same length gave a 99% significance level of 107, indicating that the observed peak is significant. The frequency of 7 corresponds to the 7 hydrophobic alpha helices in the bacteriorhodopsin structures, thus demonstrating a clear link between the Fourier transform results and structure. Application of the same method to a set of vertebrate rhodopsins also gave a significant S/N ratio, but at a peak frequency of 8. The presence of a significant S/N ratio suggests that in this case, the frequency might correspond to a genuine hydrophobicity-related property of the sequence, although it is known that the sequence has only seven distinct hydrophobic regions, corresponding to the transmembrane helices. A possible explanation of this result is the common existence of irregular length loops and of additional amino acids at the start and end of the sequence, illustrated schematically in Figure 1. These additional amino acids have the effect of increasing the frequency that is observed, as the Fourier transform method effectively fills in another peak to fit the periodicity to the hydrophobic regions that do exist. We suggest that in this case, non-periodic insertions to the periodic sequence have distorted the frequency at which periodicity is found.

When applied to the GCR2 sequences, the RRM method identified a periodicity at a frequency of seven, with a S/N value of exactly 270 (to 7 significant figures). Given a sequence length of 539, this is equal, within machine accuracy, to the maximum theoretically obtainable value for this

alignment (hence there is no need to compare to random sequences). The seven fold hydrophobicity could be interpreted as giving strong support to the idea that GCR2 is a GPCR. However, a blast search [66] of the G-protein coupled receptor sequence database (GPCRDB) did not yield any significant hits; a search of the NCBI non-redundant database yielded hits from the Lanthionine synthetase C-like protein family and a putative class B GPCR (XP_318705.3; EAA13819.3; E value 1E-45) that was probably also wrongly characterized as it also aligned well to the Lanthionine synthetase C-like protein family (results not shown). Likewise, the TMHMM and Kyte-Doolittle transmembrane helix prediction algorithms did not given any clear indication that GCR2 is a GPCR. TMPro did identify 5 of the 7 transmembrane hydrophobic regions, but TMPro only highlights transmembrane regions, it does not determine whether these are sufficiently long to span the membrane. The results of the transmembrane prediction algorithms are given as supporting information and are similar to those given elsewhere [28]. Given the negative results from the BLAST search and the transmembrane prediction algorithms, it is difficult to see why GCR2 has been proposed as a GPCR, particularly given its alignment to the Lanthionine synthetase C-like protein family. However, the origin of the confusion is apparent from the application of the RRM method to the GCR2 multiple sequence alignment. The signal to noise ratio of 270 (maximum possible = 270) compares very favourably with the signal to noise ratio of 114 (maximum possible = 115) for the bacteriorhodopsin family. Visual inspection of the GCR2 multiple sequence alignment using the hydrophobic display in jalview [41] shows 7 hydrophobic stretches which generally commence near the C-terminal end of each of the seven inner helices, they encompass the 7 GXXG motifs and end near the N-terminal region of he helix. The length of these stretches is somewhat subjective as they differs slightly for each Lanthionine synthetase C-like sequence and contains hydrophilic residues (such as His in the zinc binding site), but a conservative estimate is 15±2 and so they are generally too short to span the membrane. These regions are plotted onto the structure the lantibiotic cyclase (PDB codes 2g02) (Figure 3A) and a homology model of GCR2 created using Phyre [59] (Figure 3B). The corresponding space-filling model of 2G02 show that the exposed regions of these hydrophobic stretches map onto a single face of the protein near to the active site (Figure 3C).

Discussion

In the application of Fourier transform methods to the detection of underlying periodicities in protein sequences, some apparently useful results have been obtained [67,68]. However, care needs to be taken to ensure that such results are a product of the sequence data, rather than simply

being an artefact of the mathematics. Here we have shown that empirically-derived statistical tables can be drawn up to serve as a measure of the significance of any one given result.

The significance values for the signal-to-noise ratio derived from random sequences were much higher than might be expected. In some cases a signal-to-noise ratio of 100 (that is, a signal of 100 times greater than average magnitude) would not be significant. This can best be understood by considering what happens when large quantities of numbers are multiplied together, as occurs when the Fourier transforms are multiplied together in Equation 3.5. Where numbers between 0 and 1 are multiplied together many times, those numbers that are close to 1 remain of roughly the same magnitude, while numbers closer to zero become very small very quickly. In the case of Equation 3.5, this leads to a very high variance in the numbers produced, such that the maximal value of the sequence P(n) is much larger than the mean value, simply as an artifact of the method. Regardless of the sequence data that is fed in, very large S/N values are produced as a matter of course, and great care needs to be taken in assuming significance. This is illustrated by Figure 2, which shows mean signal-to-noise ratios for sets of 100 products of uniform random variables. This gives the equivalent of the expected signal-to-noise ratio found in the cross-spectral function (Equation 3.5) if the sequences R(n) (Equation 3.4) were 100 units long, and were distributed as uniform random variables on the interval [0,1]. From entirely random numbers, the mean signal-to-noise ratio rises above 20 for products taken across just nine sequences.

This caution about presuming levels of significance does not invalidate the method itself. Where proper care is taken to establish significance, results can be found that relate to genuine periodicities in the properties of protein sequences, as in the case of the bacteriorhodopsin set, where the seven-fold pattern in hydrophobicity reflects the seven-transmembrane helical structure of the protein. In this case also, however, care must be taken in the interpretation of results. As was demonstrated in the case of the rhodopsin sequences, factors such as insertions or deletions in the protein sequence can distort the periodicity that is found. Alignments of sequences are resistant to mutation, as long as insertions or deletions are not made in an uneven way throughout the sequence.

In another set of experiments (results not shown), a 90% rate of random residue mutation was applied to an alignment of identical, artificially constructed perfectly periodic sequences of length 300 residues (see supporting information), but the periodicity was still detectable with the RRM method, even though only 10% of the original residues remained. Similarly, random insertions made at random points in the sequence lowered the frequency at which periodicity was

found, but periodicity was still recoverable at high rates of mutation of up to 83% (here residues were removed from the end of the sequence to maintain the fixed length of 300 residues). When deletions were made from the sequence, and random residues added at the end of the sequence (again to maintain the length of 300 residues), the frequency at which periodicity was found increased, and no significant periodicity was found at a level greater than 60% mutation. Similar experiments have shown that when insertions and deletions are made in a non-regular manner then periodicity is readily destroyed. These experiments contribute towards the observation that Fourier transform methods can detect low frequency periodicity more readily than high frequency periodicity. The experiments also support the hypothesis that insertions between the helices of rhodopsin sequences can cause a distortion of the frequency at which periodicity is found, in that addition and deletion of non-periodic residues has been shown to change the frequency of periodicity. Where insertions are made between hydrophobic regions in a sequence, and the length of the sequence is *not* kept constant, the effect would be to increase the number of wavelengths that could be fitted into the sequence as a whole.

We note that periodicity as discussed here is a different concept to auto-correlation. Periodicity as discussed here implies a regular repeating pattern of residues, or of properties of residues, extending throughout the length of an entire sequence. This is a sufficient, but not necessary condition for autocorrelation, which simply measures the propensity for residues separated by a fixed-length gap to have similar properties. An example of this from mathematics would be the binary sequence $\{a_i\}$ which equals one if i is a prime number or if (i -23) is a prime number, but which equals zero otherwise. Such a sequence would have a strong autocorrelation at a distance of 23, but absolutely no periodicity. Elsewhere we will report autocorrelation in the codon and codon-pair χ^2 data even though it clearly has no statistically verifiable periodicity.

In an alternative approach to determining significance, Rackovsky [69], carrying out a Fourier-based method, compares results for protein sequences against a set of results for permutations of sequences. This gives a more accurate measure of significance than choosing a fixed value for all lengths of sequences, however, as we show elsewhere, permutations of sequences can be statistically different in nature to real protein sequences. Because of the use of multiplication to compare results, small differences between real protein sequences and their permutations can be magnified, and this has the potential to generate misleading results. In this case too, care must be taken to establish accuracy.

An illustration of a protein where statistically meaningful periodicity was identified is GCR2. Here the Fourier transform results reveal the origin of the confusion as to whether GCR2 and its lanthionine synthetase C-like homologues belong to the GPCR family. GCR2 does indeed have 7 fold hydrophobic periodicity that resides in the inner helical regions of the α-barrel and this was identified more strongly by the RRM method than the corresponding property in other well-characterized 7TM proteins such as bacteriorhodopsin and rhodopsin. These genuine hydrophobic stretches are too short to give a significant signal in most TM prediction algorithms but their presence is sufficient to yield a weak signal in some algorithms. However, the homology of GCR2 to lantibiotic cyclase [22] for which there is a crystal structure should be sufficient evidence to close the debate on the molecular nature of GCR2. Indeed, it is worth noting that Moriyama et al. used hidden Markov and related methods to identify novel plant GPCRs but they did not detect GCR2[70]. Nevertheless, while some aspects of the original report that GCR2 is the GPCR receptor for abscisic acid [21] have been seriously questioned [28], there remains the option that GCR2 may retain an indirect role in signaling in plants since not all of the experiments have been disproved in all plant tissues.

Table 1. Significance levels for signal-to-noise ratios calculated from EIIP data, derived from trials of random protein sequences. The table gives the signal-to-noise ratio that would be required for a sequence to have statistically validated genuine periodicity according to the RRM method. The signal of peak amplitude may occur at any frequency. Note that these values only apply where the protein sequences are transformed into numerical sequences according to their EIIP value. These numbers may differ if values other than EIIP are used.

	Sequence length (residues)			
	100	200	300	400
Mean value				
10 sequences	11.6	16.6	20.1	23.5
20 sequences	19.1	30.4	40.9	49.8
30 sequences	24.3	41.3	56.2	72.5
95% significance				
10 sequences	25.7	38.8	47.2	57.3
20 sequences	40.0	71.0	99.4	128.5
30 sequences	45.5	85.7	124.8	165.3
99% significance				
10 sequences	33.6	55.6	71.0	88.7
20 sequences	46.0	87.2	123.8	167.7
30 sequences	48.8	95.9	140.7	190.1

Figure Legends

Fig. 1. Location of helices (marked as black blocks along the sequence) in bacteriorhodopsins and vertebrate rhodopsins, (sequences from the GPCRDB). The bacteriorhodopsins (top) have 7 essentially equally-spaced hydrophobic helices, leading to a significant spike related to hydrophobicity score at a frequency of 7 in the cross-spectral function. In the rhodopsin sequences (below), extended non-hydrophobic regions, e.g. between helices 4 and 5 lead to a peak in the cross-spectral function at a frequency of 8.

Fig. 2. Mean signal-to-noise ratios calculated from 100 000 sets of 100 products of t U[0,1] random variables.

Fig. 3. (A) The structure of 2G02, with the 7 hydrophobic regions mapped onto the 7 inner helices in shown black (or various shades of green online), that contain the 7 GXXG motifs (cyan online); the key residues of the active site are displayed in space-filling mode. **(B)** The structure of GCR2, with the 7 hydrophobic regions mapped onto the 7 inner helices shown in black (or various shades of green online) that contain the 7 GXXG motifs (cyan online); the key residues of the active site are displayed in space-filling mode. Residues 255-260 are omitted. **(C)** The structure of 2G02, shown in space-filling mode, indicating that the 7 hydrophobic regions, shown in black (or various shades of green online), map onto a single surface. The 7 GXXG motifs are shown in cyan in the online version and the key residues of the active site are displayed in space-filling mode.

References

1, J.M.Zimmerman, N.Eliezer and R.Simha,The characterization of amino acid sequences in proteins by statistical methods, J Theor.Biol. 21 (1968) 170-201.

2, A.D.McLachlan and M.Stewart,The 14-fold periodicity in alpha-tropomyosin and the interaction with actin, J Mol.Biol. 103 (1976) 271-298.

3, G.N.Phillips, Jr., J.P.Fillers and C.Cohen,Tropomyosin crystal structure and muscle regulation, J Mol.Biol. 192 (1986) 111-131.

4, G.N.Phillips, Jr.,Construction of an atomic model for tropomyosin and implications for interactions with actin, J Mol.Biol. 192 (1986) 128-131.

5, V.R.Chechetkin and A.Y.Turygin,Search of hidden periodicities in DNA sequences, J Theor.Biol. 175 (1995) 477-494.

6, V.R.Chechetkin and V.V.Lobzin,Nucleosome units and hidden periodicities in DNA sequences, J Biomol.Struct.Dyn. 15 (1998) 937-947.

7, V.V.Lobzin and V.R.Chechetkin,Order and correlations in genomic DNA sequences. The spectral approach, Uspekhi Fizicheskikh Nauk 170 (2000) 57-81.

8, B.D.Silverman and R.Linsker,A measure of DNA periodicity, J Theor.Biol. 118 (1986) 295-300.

9, J.L.Cornette, K.B.Cease, H.Margalit, J.L.Spouge, J.A.Berzofsky and C.DeLisi,Hydrophobicity scales and computational techniques for detecting amphipathic structures in proteins, J Mol.Biol. 195 (1987) 659-685.

10, D.Donnelly, M.S.Johnson, T.L.Blundell and J.Saunders,An analysis of the periodicity of conserved residues in sequence alignments of G-protein coupled receptors. Implications for the three-dimensional structure, FEBS Lett. 251 (1989) 109-116.

11, D.Donnelly, J.P.Overington, S.V.Ruffle, J.H.Nugent and T.L.Blundell,Modeling alpha-helical transmembrane domains: the calculation and use of substitution tables for lipid-facing residues, Protein Sci. 2 (1993) 55-70.

12, D.Donnelly, J.P.Overington and T.L.Blundell,The prediction and orientation of alpha-helices from sequence alignments: the combined use of environment-dependent substitution tables, Fourier transform methods and helix capping rules, Protein Eng 7 (1994) 645-653.

13, D.Eisenberg, R.M.Weiss and T.C.Terwilliger,The hydrophobic moment detects periodicity in protein hydrophobicity, Proc.Natl.Acad.Sci.U.S.A 81 (1984) 140-144.

14, H.Leonov and I.T.Arkin,A periodicity analysis of transmembrane helices, Bioinformatics 21 (2005) 2604-2610.

15, E.V.Korotkov, M.A.Korotkova, V.M.Rudenko and K.G.Skriabin,[Regions with latent periodicity in protein amino acid sequences], Mol.Biol.(Mosk) 33 (1999) 688-695.

16, V.J.Makeev and V.G.Tumanyan,Search of periodicities in primary structure of biopolymers: a general Fourier approach, Comput.Appl.Biosci. 12 (1996) 49-54.

17, E.V.Korotkov, M.A.Korotkova, F.E.Frenkel' and N.A.Kudriashov,[The informational concept of searching for periodicity in symbol sequences], Mol.Biol.(Mosk) 37 (2003) 436-451.

18, V.P.Turutina, A.A.Laskin, N.A.Kudryashov, K.G.Skryabin and E.V.Korotkov,Identification of amino acid latent periodicity within 94 protein families, J Comput.Biol. 13 (2006) 946-964.
19, S.Rackovsky,"Hidden" sequence periodicities and protein architecture, Proc.Natl.Acad.Sci.U.S.A 95 (1998) 8580-8584.
20, C.H.de Trad, Q.Fang and I.Cosic,Protein sequence comparison based on the wavelet transform approach, Protein Eng 15 (2002) 193-203.
21, X.G.Liu, Y.L.Yue, B.Li, Y.L.Nie, W.Li, W.H.Wu and L.G.Ma,A G protein-coupled receptor is a plasma membrane receptor for the plant hormone abscisic acid, Science 315 (2007) 1712-1716.
22, B.Li, J.P.Yu, J.S.Brunzelle, G.N.Moll, W.A.van der Donk and S.K.Nair,Structure and mechanism of the lantibiotic cyclase involved in nisin biosynthesis, Science 311 (2006) 1464-1467.
23, H.Mayer, H.Bauer and R.Prohaska,Organization and chromosomal localization of the human and mouse genes coding for LanC-like protein 1 (LANCL1), Cytogenet.Cell Genet. 93 (2001) 100-104.
24, H.Bauer, H.Mayer, A.Marchler-Bauer, U.Salzer and R.Prohaska,Characterization of p40/GPR69A as a peripheral membrane protein related to the lantibiotic synthetase component C, Biochem.Biophys.Res.Commun. 275 (2000) 69-74.
25, T.Hirayama and K.Shinozaki,Perception and transduction of abscisic acid signals: keys to the function of the versatile plant hormone ABA, Trends Plant Sci. 12 (2007) 343-351.
26, H.Mayer, J.Breuss, S.Ziegler and R.Prohaska,Molecular characterization and tissue-specific expression of a murine putative G-protein-coupled receptor, Biochim.Biophys.Acta 1399 (1998) 51-56.
27, H.Mayer, U.Salzer, J.Breuss, S.Ziegler, A.Marchler-Bauer and R.Prohaska,Isolation, molecular characterization, and tissue-specific expression of a novel putative G protein-coupled receptor, Biochim.Biophys.Acta 1395 (1998) 301-308.
28, Y.Gao, Q.Zeng, J.Guo, J.Cheng, B.E.Ellis and J.G.Chen,Genetic characterization reveals no role for the reported ABA receptor, GCR2, in ABA control of seed germination and early seedling development in Arabidopsis, Plant J (2007).
29, C.H.de Trad, Q.Fang and I.Cosic,Protein sequence comparison based on the wavelet transform approach, Protein Eng 15 (2002) 193-203.
30, C.H.Wu, R.Apweiler, A.Bairoch, D.A.Natale, W.C.Barker, B.Boeckmann, S.Ferro, E.Gasteiger, H.Huang, R.Lopez, M.Magrane, M.J.Martin, R.Mazumder, C.O'Donovan, N.Redaschi and B.Suzek,The Universal Protein Resource (UniProt): an expanding universe of protein information, Nucleic Acids Res. 34 (2006) D187-D191.
31, I.Cosic,Macromolecular bioactivity: is it resonant interaction between macromolecules?--Theory and applications, IEEE Trans.Biomed.Eng 41 (1994) 1101-1114.
32, I.Cosic and D.Nesic,Prediction of 'hot spots' in SV40 enhancer and relation with experimental data, Eur.J.Biochem. 170 (1987) 247-252.
33, V.Veljkovic and I.Slavic,Simple General Model Pseudopotential, Physical Review Letters 29 (1972) 105-&.
34, V.Veljkovic and D.I.Lalovic,Correlation between the carcinogenicity of organic substances and their spectral characteristics, Experientia 34 (1978) 1342-1343.
35, V.Veljkovic, A Theoretical Approach to the Preselection of Carcinogens and Chemical Carcinogenesis (Gordon and Breach Science Publishers, 1980).
36, V.Veljkovic, I.Cosic, B.Dimitrijevic and D.Lalovic,Is it possible to analyze DNA and protein sequences by the methods of digital signal processing?, IEEE Trans.Biomed.Eng 32 (1985) 337-341.

37, V.J.Veljkovic and D.I.Lalovic,Theoretical prediction of mutagenicity and carcinogenicity of chemical substances, Cancer Biochem.Biophys. 1 (1976) 295-298.
38, W.S.Barnes and D.E.Levin,Theoretical prediction of carcinogenicity: quasi-quantification by quasi-valence, Experientia 35 (1979) 565-567.
39, L.Pauling,Book review: A Theoretical Approach to the Preselection of Carcinogens and Chemical Carcinogenesis. by Veljko Veljkovic, The Quarterly Review of Biology 57 (1982) 228-229.
40, F.Horn, J.Weare, M.W.Beukers, S.Horsch, A.Bairoch, W.Chen, O.Edvardsen, F.Campagne and G.Vriend,GPCRDB: an information system for G protein-coupled receptors, Nucleic Acids Res. 26 (1998) 275-279.
41, M.Clamp, J.Cuff, S.M.Searle and G.J.Barton,The Jalview Java alignment editor, Bioinformatics 20 (2004) 426-427.
42, T.Kulikova, R.Akhtar, P.Aldebert, N.Althorpe, M.Andersson, A.Baldwin, K.Bates, S.Bhattacharyya, L.Bower, P.Browne, M.Castro, G.Cochrane, K.Duggan, R.Eberhardt, N.Faruque, G.Hoad, C.Kanz, C.Lee, R.Leinonen, Q.Lin, V.Lombard, R.Lopez, D.Lorenc, H.McWilliam, G.Mukherjee, F.Nardone, M.Pilar, G.Pastor, S.Plaister, S.Sobhany, P.Stoehr, R.Vaughan, D.Wu, W.M.Zhu and R.Apweiler,EMBL Nucleotide Sequence Database in 2006, Nucleic Acids Research 35 (2007) D16-D20.
43, G.A.Gutman and G.W.Hatfield,Nonrandom utilization of codon pairs in Escherichia coli, Proc.Natl.Acad.Sci.U.S.A 86 (1989) 3699-3703.
44, J.R.Buchan, L.S.Aucott and I.Stansfield,tRNA properties help shape codon pair preferences in open reading frames, Nucleic Acids Res. 34 (2006) 1015-1027.
45, B.Irwin, J.D.Heck and G.W.Hatfield,Codon pair utilization biases influence translational elongation step times, J.Biol.Chem. 270 (1995) 22801-22806.
46, F.Horn, J.Weare, M.W.Beukers, S.Horsch, A.Bairoch, W.Chen, O.Edvardsen, F.Campagne and G.Vriend,GPCRDB: an information system for G protein-coupled receptors, Nucleic Acids Res. 26 (1998) 275-279.
47, S.H.White and W.C.Wimley,Hydrophobic interactions of peptides with membrane interfaces, Biochim.Biophys.Acta 1376 (1998) 339-352.
48, I.Cosic,Macromolecular bioactivity: is it resonant interaction between macromolecules?--Theory and applications, IEEE Trans.Biomed.Eng 41 (1994) 1101-1114.
49, I.Cosic,Macromolecular bioactivity: is it resonant interaction between macromolecules?--Theory and applications, IEEE Trans.Biomed.Eng 41 (1994) 1101-1114.
50, I.Cosic,Macromolecular bioactivity: is it resonant interaction between macromolecules?--Theory and applications, IEEE Trans.Biomed.Eng 41 (1994) 1101-1114.
51, E.L.Sonnhammer, H.G.Von and A.Krogh,A hidden Markov model for predicting transmembrane helices in protein sequences, Proc.Int.Conf.Intell.Syst.Mol.Biol. 6 (1998) 175-182.
52, A.Krogh, B.Larsson, H.G.Von and E.L.Sonnhammer,Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes, J.Mol.Biol. 305 (2001) 567-580.
53, J.Kyte and R.F.Doolittle,A simple method for displaying the hydropathic character of a protein, J.Mol.Biol. 157 (1982) 105-132.
54, A.Bateman, E.Birney, R.Durbin, S.R.Eddy, R.D.Finn and E.L.Sonnhammer,Pfam 3.1: 1313 multiple alignments and profile HMMs match the majority of proteins, Nucleic Acids Res. 27 (1999) 260-262.
55, J.D.Thompson, D.G.Higgins and T.J.Gibson,CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice, Nucleic Acids Res. 22 (1994) 4673-4680.

56, J.D.Thompson, T.J.Gibson, F.Plewniak, F.Jeanmougin and D.G.Higgins,The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools, Nucleic Acids Res. 25 (1997) 4876-4882.
57, J.Shi, T.L.Blundell and K.Mizuguchi,FUGUE: sequence-structure homology recognition using environment-specific substitution tables and structure-dependent gap penalties, J Mol.Biol. 310 (2001) 243-257.
58, D.T.Jones,GenTHREADER: an efficient and reliable protein fold recognition method for genomic sequences, J Mol.Biol. 287 (1999) 797-815.
59, R.M.nett-Lovsey, A.D.Herbert, M.J.Sternberg and L.A.Kelley,Exploring the extremes of sequence/structure space with ensemble fold recognition in the program Phyre, Proteins (2007).
60, V.Veljkovic, I.Cosic, B.Dimitrijevic and D.Lalovic,Is it possible to analyze DNA and protein sequences by the methods of digital signal processing?, IEEE Trans.Biomed.Eng 32 (1985) 337-341.
61, I.Cosic, A.N.Hodder, M.I.Aguilar and M.T.Hearn,Resonant recognition model and protein topography. Model studies with myoglobin, hemoglobin and lysozyme, Eur.J.Biochem. 198 (1991) 113-119.
62, I.Cosic,Macromolecular bioactivity: is it resonant interaction between macromolecules?--Theory and applications, IEEE Trans.Biomed.Eng 41 (1994) 1101-1114.
63, I.Cosic,Virtual spectroscopy for fun and profit, Biotechnology (N.Y.) 13 (1995) 236-238.
64, I.Cosic and E.Pirogova,Bioactive peptide design using the Resonant Recognition Model, Nonlinear.Biomed.Phys. 1 (2007) 7.
65, C.H.de Trad, Q.Fang and I.Cosic,Protein sequence comparison based on the wavelet transform approach, Protein Eng 15 (2002) 193-203.
66, S.F.Altschul, T.L.Madden, A.A.Schaffer, J.Zhang, Z.Zhang, W.Miller and D.J.Lipman,Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, Nucleic Acids Res. 25 (1997) 3389-3402.
67, C.H.de Trad, Q.Fang and I.Cosic,Protein sequence comparison based on the wavelet transform approach, Protein Eng 15 (2002) 193-203.
68, S.Rackovsky,"Hidden" sequence periodicities and protein architecture, Proc.Natl.Acad.Sci.U.S.A 95 (1998) 8580-8584.
69, S.Rackovsky,"Hidden" sequence periodicities and protein architecture, Proc.Natl.Acad.Sci.U.S.A 95 (1998) 8580-8584.
70, E.N.Moriyama, P.K.Strope, S.O.Opiyo, Z.Chen and A.M.Jones,Mining the Arabidopsis thaliana genome for highly-divergent seven transmembrane receptors, Genome Biol. 7 (2006) R96.

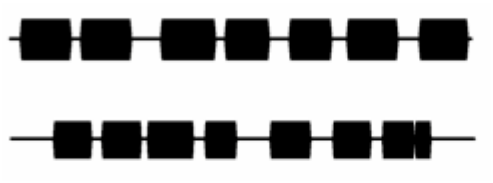


Fig 1.

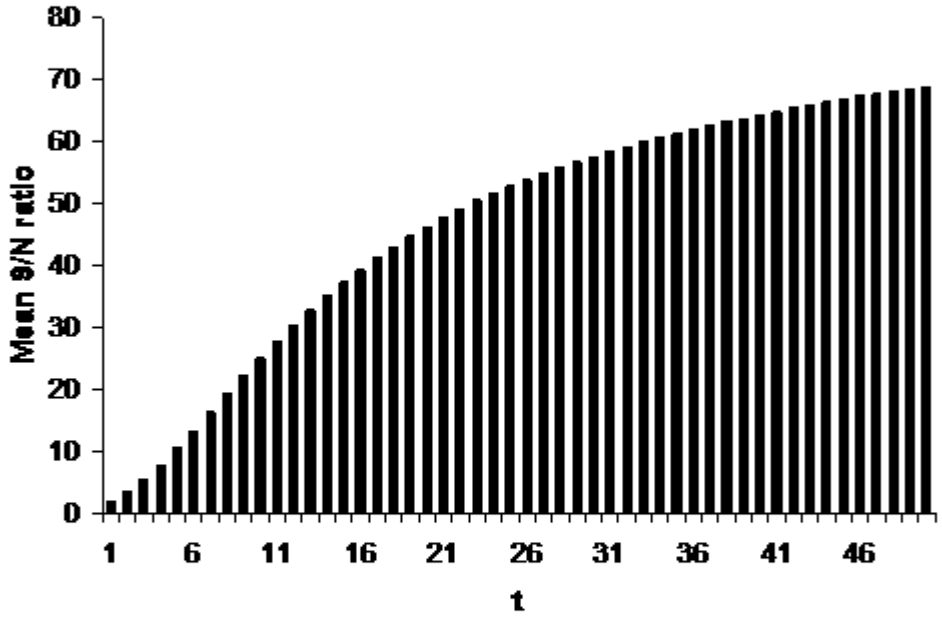


Fig. 2.

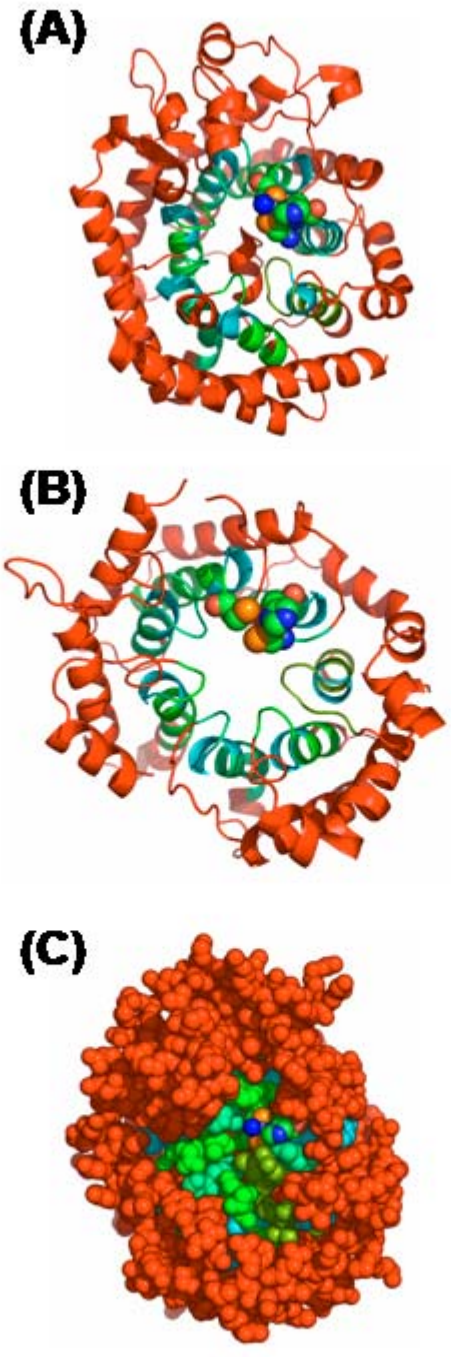


Fig. 3_