Identification of novel multi-transmembrane proteins from genomic databases using quasi-periodic structural properties

Junhyong Kim¹, ², ³, ⁴, Etsuko N. Moriyama¹, Coral G. Warr³, Peter J. Clyne³ and John R. Carlson³

¹ Department of Ecology and Evolutionary Biology, ² Department of Statistics, and ³ Department of Molecular, Cellular and Developmental Biology, Yale University, USA

Abstract

Motivation: Identification of novel G protein-coupled receptors and other multi-transmembrane proteins from genomic databases using structural features.

Results: Here we describe a new algorithm for identifying multi-transmembrane proteins from genomic databases with a specific application to identifying G protein-coupled receptors (GPCRs) that we call quasi-periodic feature classifier (QFC). The QFC algorithm uses concise statistical variables as the ‘feature space’ to characterize the quasi-periodic physico-chemical properties of multi-transmembrane proteins. For the case of identifying GPCRs, the variables are then used in a non-parametric linear discriminant function to separate GPCRs from non-GPCRs. The algorithm runs in time linearly proportional to the number of sequences, and performance on a test dataset shows 96% positive identification of known GPCRs. The QFC algorithm also works well with short random segments of proteins and it positively identified GPCRs at a level greater than 90% even with segments as short as 100 amino acids. The primary advantage of the algorithm is that it does not directly use primary sequence patterns which may be subject to sampling bias. The utility of the new algorithm has been demonstrated by the isolation from the Drosophila genome project database of a novel class of seven-transmembrane proteins which were shown to be the elusive olfactory receptor genes of Drosophila.

Introduction

Membrane proteins govern the transport and signaling of specific molecules and act as gateways between the cell (or organelle) and its environment. Due to this role, many membrane proteins play a critical part in the normal physiological function of the organism—they transport specific metabolites, drugs, and ions and they act as receptors for a wide variety of signaling molecules, a few of which include hormones, nucleotides, opiates, neurotransmitters, and odorants. (e.g. Beja and Bibi, 1995; Dewji and Singer, 1997; Hildebrand, 1997; Lancet et al., 1993; von Heijne and Manoil, 1990). Membrane proteins are important targets of pharmacological agents, and identifying and classifying the functions of them is a significant problem for both basic and applied research. In this paper we describe a new algorithm for identifying multi-transmembrane proteins such as G Protein-Coupled Receptors (GPCR) from large-scale genomic databases.

Existing methods of identifying proteins rely mostly on putative local or global sequence homology and the principle of conservation (e.g. Attwood et al., 1998; Attwood and Findlay, 1993; Bairoch, 1992; Baldi and Chauvin, 1994; Barton and Sternberg, 1990; Kulp et al., 1997; Sander and Schneider, 1991). Although this approach is tremendously useful, there are two potential problems in applications to a large protein class such as, say, the GPCR membrane proteins. Firstly, the proteins may have diverged so much that little sequence similarity remains from which to infer homology. Secondly and more importantly, a class of proteins is usually established by some putative conserved motifs and domains. Typically such motifs and domains are initially identified from a few examples. Since subsequent novel proteins are assigned to classes and families through the same motifs, possible initial sampling bias is reinforced. An example of this is that until recently, no insect odorant receptors had been identified despite numerous attempts at cloning through

© Oxford University Press 2000
homology principles (Clyne et al., 1999; Vosshall et al., 1999). Avoiding sampling bias requires identification of protein classes by more global (and possibly more functional) properties than primary amino acid sequence patterns. In fact, a large volume of literature describes the use of amino acid physico-chemical information to infer protein structure (e.g. Aloy et al., 1997; Casadio et al., 1996; Cserzo et al., 1997; Edelman, 1993; Juretic et al., 1998; Kuhn and Leigh, 1985; Rost et al., 1995; Rost and O’Donoghue, 1997; Weiss et al., 1993). However, the focus of these studies has mainly been structure prediction (especially the transmembrane alpha helix) and these structural inferences have not been commonly applied to large-scale genomic data analysis. Huang et al. (1996) used a transmembrane structure prediction neural network for large-scale screens but the research did not use the information in multi-transmembrane domains and it also had to sacrifice accurate prediction for computational efficiency. The two main difficulties here are: firstly, the computational problem associated with complex models; and secondly, the problem of converting descriptive information such as amino-acid hydropathy plots into statistically meaningful information for pattern recognition and classification.

Here, we describe a new algorithm that overcomes these difficulties and that was used to identify a novel class of odorant receptors and taste receptors in Drosophila (Clyne et al., 1999). We call this new algorithm quasi-periodic feature classifier (QFC). Recently, Jaakkola et al. (1999) noted the difference in gene identification by generative methods such as hidden Markov models and classification methods. In the latter, sequences are placed in a suitable ‘feature space’ and discrimination functions are constructed that classify the sequences into specific categories (e.g. GPCRs and non-GPCRs). The ‘feature space’ + discrimination function approach has the advantage of being computationally efficient and more easily generalized to remotely related sequences. In these terms, we construct a ‘feature space’ using statistical measures of amino-acid physico-chemical properties and then use a linear discriminant function to extract GPCR from genomic databases. The key to this approach is constructing the most useful feature space. For multi-transmembrane proteins, we constructed a feature space by using the information in the quasi-periodic interplay between the local structure (transmembrane alpha helix) and global structure (repeated multiple domains) and characterizing this information with concise statistical variables. Here we describe the QFC algorithm, show its performance with test data, and then illustrate its utility in the identification of Drosophila olfactory genes.

![Flow chart of the procedures involved in the algorithm development and calculation. See text for details.](image)

**The QFC algorithm**

Our strategy was to statistically characterize the differentiating features of the physico-chemical properties of known multi-transmembrane proteins using heuristic data reduction principles (Tukey, 1977); that is, reducing noise in the data and extracting sufficient statistics to allow multi-transmembrane proteins to be distinguished from all other proteins. The parameters derived from this characterization were then used to screen databases for novel proteins. Each step of this strategy is outlined by a flow chart in Figure 1 and described in detail in the following experimental procedures.

**Training data sets and test data sets**

Our algorithm depends on extracting classification variables that separate multi-transmembrane proteins from all others. Our initial goal was to use the algorithm to distinguish GPCRs from non-GPCRs. Therefore the classification variables were extracted based on a ‘training data set’ which was compiled from a list of putative GPCRs in the GPCR database (http://swift.embl-heidelberg.de/7tm) and from random SWISSPROT proteins. A total of 750 GPCRs and 1000 random non-GPCR proteins, between 200 amino acids and 1000 amino acids in length, were selected as the training data set. Independent of the training data set, 100 putative GPCRs and 100 random
Discriminant function detection of membrane proteins

Table 1. A list of coefficients used for the amino-acid usage index. The coefficients were multiplied by the amino-acid frequency of a protein and summed, yielding a linear function that best discriminates GPCRs from non-GPCRs

<table>
<thead>
<tr>
<th>K</th>
<th>N</th>
<th>T</th>
<th>R</th>
<th>S</th>
</tr>
</thead>
<tbody>
<tr>
<td>36.7455</td>
<td>60.6998</td>
<td>15.974</td>
<td>-2.63569</td>
<td>-16.3077</td>
</tr>
<tr>
<td>10.8843</td>
<td>63.7699</td>
<td>82.7947</td>
<td>63.0038</td>
<td>36.2475</td>
</tr>
<tr>
<td>53.0168</td>
<td>120.547</td>
<td>6.66418</td>
<td>75.3693</td>
<td>-8.97487</td>
</tr>
<tr>
<td>90.9217</td>
<td>-13.1554</td>
<td>-143.336</td>
<td>-17.4348</td>
<td>0.0</td>
</tr>
</tbody>
</table>

Fig. 2. The performance of the new Quasi-periodic Feature Classifier (QFC) algorithm on short protein fragments. The y-axis shows percent of input sequences identified as GPCRs. The x-axis shows the fragment size. The solid lines indicate the average % of GPCRs positively identified by different algorithms. The dashed-lines show the fragment size. The solid lines indicate the average % of GPCRs positively identified by the different algorithms. (Some methods yielded no false positives.) The symbols for each algorithm are: open squares (QFC), solid triangle (PROSITE profile using Pfscan), solid diamond (PROSITE pattern using regular expression search), solid circles (Pfam profile HMM), and ‘+’s (PRINTS finger print search), respectively.

non-GPCR proteins were reserved for performance tests of the algorithm.

Sliding window recognizer and amino-acid usage bias

Firstly the physico-chemical properties of the amino acids in the molecules were characterized using various indices and standard measurements. The literature describes many different possible measures including indices of hydropathy (Engelman et al., 1986; Kyte and Doolittle, 1982), polarity, pl, pKa, molecular weight, amino acid composition, etc. (Many of these properties are correlated with each other.) A protein sequence is therefore described by a set of variables, \( x_1 - x_n \) and for each \( x_i \), we have a value \( x_{ij} \) for the \( i \)th amino-acid index value for the \( j \)th position.

Therefore, \( x_{ij} - x_{jk} \) constitutes a profile of the protein in terms of the \( i \)th amino-acid property index. In this study, the use of GES hydropathy index (Engelman et al., 1986), Kyte-Doolittle index (Kyte and Doolittle, 1982), polarity, pl, molecular weight, solubility (Brown, 1998), and alpha helix index (Deleage and Roux, 1987) were examined. We
also examined the utility of each measure for differentiating GPCR and non-GPCR proteins in our data set through preliminary exploratory analysis described below.

The raw profile $x_{ij} - x_{ik}$ is a very noisy characterization of the protein. The noise was reduced with the widely used Sliding Window Recognizer (SWR von Heijne, 1992, 1994) where the for each $j$th position we have,

$$x_{ij} = \sum_{k=-d}^{d} w_{j-k} x_{j-k}$$

The particular kernel window function, $w$, of the sliding summary is critical for recognizing features at the proper scale. For example, a small window will be sensitive to small fluctuations in the physico-chemical measurements while a large window of, say 100 amino acids, will only summarize the overall trend. The kernel function represents the first significant data reduction. For our data, we used a 15 amino acid Gaussian kernel convoluted with a 16 amino acid constant kernel which worked well to characterize the typical period of transmembrane domains ($\sim$16 to 22 amino acids; we also tested 32 and 64 amino acid Gaussian kernels, which did not capture the periodic local structure as well).

Next we wished to characterize the profile with a compact set of descriptors (the second data reduction). While more sophisticated waveform analyses such as the Fourier transform are available, we chose to produce as simple a characterization as possible. To be useful for short sequences, an important criterion for this purpose was that the descriptors should be as local as possible (e.g. the maximum value of a profile is a global property of the protein and not a local property). We examined the use of periodicity (defined below), variance of periodicity, first-order derivative, second-order derivative, and variance of the derivative. Through exploratory analysis, average periodicity and the variance of the first-order derivative were chosen as the optimal statistical variables characterizing each profile. The precise definitions of these variables are as follows:

**Average periodicity.** Is a characterization of how frequently the sliding window profile crosses over a neutral value. The neutral values were $-0.5$ and $8.325$ for the GES scale and polarity, respectively (Engelman et al., 1986). Average periodicity is computed by simply counting how many times the profile changes sign around the neutral value and then dividing by the length of the sequence. For better discrimination, the average periodicity was log-transformed.

**Variance of the 1st order derivative.** Since the positions along the protein are discrete, we define the 1st order derivative of the $i$th profile at position $j$ as $x_{ij} - x_{ij-1}$. Variance of the 1st order derivative is the variance of these values computed over the entire protein sequence.

Next, we also attempted to extract an index of amino-acid usage bias that is typical of GPCRs (see, for example, Wootton, 1994). We computed an amino-acid usage index by counting the frequency of each amino acid in GPCRs (the whole protein) and non-GPCRs in our training data set and constructing a linear discriminant function (Gnanadesikan, 1977) of the amino acid frequencies that maximally separated the two classes of proteins (see below for more details). If $f_i$ is the frequency of the $i$th amino-acid (out of the 20 possible) in the protein, the linear discriminant function for amino-acid frequencies is a function,

$$L_{AAF} = \sum_{i=1}^{20} c_i f_i$$

such that the two classes of proteins are best separated by the function value, $L_{AAF}$ (see also below). The coefficients, $c_i$, derived from the discriminant analysis are shown in Table 1.

In total, seven different profiles characterized by two statistics (average periodicity and variance of 1st order derivative) were examined and their utility assessed in discriminating GPCRs from non-GPCRs by a step-wise deletion procedure. That is, all 14 variables were first used for the discrimination, then each variable was deleted in turn and posterior performance was assessed by the linear discriminant function described in the next section. One of our goals was to reduce the number of spurious variables which may result in over-fitting the data. Based on the step-wise deletion analysis, we settled on the following four variables:

1. Amino-acid usage index.
2. Log of the average periodicity of the GES scale.
3. Log of the average periodicity of the polarity scale.
4. Variance of the first derivative of the polarity scale.

**Classification of GPCR and non-GPCR proteins**

Once the amino acid sequences are characterized by our set of statistical variables the next step is to classify them into the desired protein classes (i.e. GPCR versus non-GPCR).

### Table 2. The final coefficients for the non-parametric linear discriminant function. Four variables are used in the linear discriminant function with the coefficients shown in this table.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td>Amino-acid usage index</td>
<td>$-1.670$</td>
</tr>
<tr>
<td>Log average GES periodicity</td>
<td>$-3.334$</td>
</tr>
<tr>
<td>Log average polarity periodicity</td>
<td>$19.485$</td>
</tr>
<tr>
<td>Variance of first order derivative</td>
<td>$11.237$</td>
</tr>
</tbody>
</table>
non-GPCR). As our classification method we used a non-parametric variant of discriminant function analysis (Gnanadesikan, 1977). The idea of a DFA is to obtain a set of functions of the original variables (e.g. average periodicity of the polarity scale) such that when the observations are scored by the function values, the resulting projection maximally separates some pre-defined classes. Once such functions are derived, they can be applied to novel data. Here, we used a simple linear function since a linear function is sufficient to separate two canonical classes when the distribution of the objects in each class is approximately convex (which is the case for our data),

\[ L_{QFC} = \sum_{i=1}^{4} c_i s_i \]

where \( s_i \) is the \( i \)th sliding window statistic and \( c_i \) is the weight coefficient.

The parametric equations commonly used to derive the best linear discriminant axes rely on statistical distribution assumptions that are not applicable to our data. Therefore, we used numerical optimizations of a non-parametric classification criterion to estimate the best coefficients of the linear discriminant function. The non-parametric criterion was the ‘runs’ criterion where a run is defined as a contiguous series of GPCRs (or non-GPCRs) along the linear discriminant function. Hypothetically, if the two classes of proteins are exactly separated we would have two runs, a run of GPCRs followed by a run of non-GPCRs. We used a numerical optimization algorithm to minimize the number of alternating runs of GPCR and non-GPCR proteins along the discriminant axis.

The optimal discrimination coefficients were derived from the training data and these values are shown in Table 2. The optimal cutoff values depend on the desired level of false positives and false negatives. For the training data, we used the criteria \(-41.9 < L_{QFC} < \infty\) for classifying the candidate sequence as GPCR for the result shown in Table 3. The histogram of the training data set on the linear discriminant function can also be used to derive an empirical probability distribution for GPCR or non-GPCR classification. Alternatively, a ranking of the candidates can be derived from the distance of the candidate sequences (on the discriminant axis) to the centroid of the training set distribution.

### QFC algorithm performance

The QFC algorithm runs in time linearly proportional to the number of sequences and therefore it is trivial to screen very large databases. For example, it took \(~\sim 1\) min. to process \(~\sim 30\) million bases of \(Drosophila\) data on a 400 MHz Pentium II machine. The discriminability of the algorithm was tested on an independent test data set of 100 GPCRs and 100 non-GPCRs. The QFC algorithm was initially tested on full length sequences and then on short random subsequences to see how it might perform on EST databases or open reading frames broken up by introns.

For comparative purposes we also attempted to identify GPCRs by using four other existing methods. PROSITE pattern search was carried out with the GPCR motifs (PD0C00210, PD0C00559, PD0C00754 Bairoch, 1992) converted to the equivalent regular expressions. Two profile searches were performed. The GPCR profiles (PS50261, PS50262, PS50259) taken from PROSITE generalized profile library (Bucher et al., 1996) were searched by the program ‘pfscan’ (a component of ‘pfutils 2.2’ package available from http://www.isrec.isb-sib.ch/ftp-server/pfutils/pft2.2/).

There are four GPCR entries (7tm1, 7tm2, 7tm3, 7tm4) in Pfam 4.0 profile HMM (Hidden Markov Model) database (Durbin et al., 1998). The searches against these

---

**Table 3.** A comparison among different search methods. The training data set includes 750 GPCRs and 1000 random non-GPCR proteins. The test data set includes 100 GPCRs and 100 random non-GPCR proteins. The performance of the five methods are compared for the percentage of positive identification of GPCRs (positive %) and the percentage of false positive identification (false positive %). For PROSITE profile search, Pfam profile HMM search, and PRINTS fingerprint search, the default cut-off value given by each search program was used. For the PRINTS search, the ‘‘2TBH’ line-tag in the output was looked for to identify the match with the score higher than the cut-off value. All searches were performed against the subsets of databases that contain only GPCR-related entries (see the text for those IDs).

<table>
<thead>
<tr>
<th>Method</th>
<th>Training data set</th>
<th>Test data set</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Positive %</td>
<td>False positive %</td>
</tr>
<tr>
<td></td>
<td>(no. of hits)</td>
<td>(no. of hits)</td>
</tr>
<tr>
<td>New QFC algorithm</td>
<td>97 (727)</td>
<td>2.1 (21)</td>
</tr>
<tr>
<td>PROSITE pattern</td>
<td>93 (701)</td>
<td>0 (1)</td>
</tr>
<tr>
<td>PROSITE profile using pfscan</td>
<td>99 (741)</td>
<td>0 (0)</td>
</tr>
<tr>
<td>Pfam profile HMMs</td>
<td>98 (738)</td>
<td>0 (1)</td>
</tr>
<tr>
<td>PRINTS</td>
<td>99 (744)</td>
<td>0 (3)</td>
</tr>
</tbody>
</table>
profiles were performed by the program ‘hmmpfam’ (a part of ‘HMMER 2.1.1’ http://hmmer.wustl.edu). Another search was performed on PRINTS fingerprint database (Attwood et al., 1999). The ‘FingerPRINTScan’ program (available from ftp://ftp.biochem.ucl.ac.uk/pub/scordis/fignerPRINTScan/) was used to search against 172 GPCR fingerprints taken from PRINTS 22.0.

On the training data set, the QFC algorithm provided excellent discrimination of GPCRs and non-GPCRs. Approximately 97% of the GPCRs in the training data set were correctly identified while only 2.1% of the non-GPCRs were mis-identified as GPCR (false positives) using the \(< - 41.9\) cutoff value. On the independent test data set of 100 GPCRs and 100 non-GPCRs the results of our algorithm were 99% positive identification and 0% false-positives. In Table 3, we compared the QFC algorithm with the four methods. Our QFC algorithm, the two profile-based methods (PROSITE profile search and Pfam profile HMM search), and the PRINTS fingerprint search performed equally well on the test dataset.

The above performances were measured on full-length proteins. In practice, genomic coding sequences are interrupted by introns, and EST databases consist of fragments of proteins typically ranging in size from 100 to 150 amino acids. Therefore, one of our objectives was for our algorithm to perform well on short fragments of proteins. To test for this kind of performance, we created a protein fragment test data set by taking the 100 test sequences and randomly choosing smaller subsequences from each protein. The sizes of the subsequences chosen were 50, 75, 100, 150, 200, 250, 300, and 400 amino acids. For each subsequence size, 50 random replicate subsequences were produced. Figure 2 shows the results of applying the QFC algorithm vs those from the other four methods. The plot shows the average percentage of GPCRs identified (over the 50 replicates) as a function of amino-acid length. As can be seen, the QFC algorithm greatly outperforms the other methods. Even with 50 amino acids our algorithm obtains \(\sim 78\%\) positive identification, whereas at this length Pfam profile HMM search could identify less than 50% of GPCRs correctly. The performance of the other three methods was even lower. At the typical range of sizes for sequences in EST databases (100–150 amino acids), our algorithm shows better than 90% identification.

In order to further examine how specifically the QFC algorithm can discriminate GPCR sequences from other multi-transmembrane protein sequences, 530 non-GPCR ‘ionic channel’ entries from SWISSPROT 38.0 were collected (‘ion channels’ are a very heterogeneous family and we did not attempt to carefully curate the 530 sequence database). The algorithm was retrained to discriminate GPCR vs ion channels. After exploratory analysis as described above, we found best discrimination using four variables, \(L_{\text{aaf}}, \text{log average GES periodicity, variance of 1st order molecular weight derivative, and derivative. We obtained 96.4\% positive identification of GPCRs from ion channels with 2.0\% false positive rate. We next tested the ability of the algorithm to discriminate ion channels from random proteins. We again retrained the algorithm using the same four variables and obtained 90\% positive identification of ion channels with 11.8\% false positive rate. As mentioned ion channels are an extremely heterogeneous family and we expect better curation, sub-classification, and training to lead to better performance.

Lastly, we examined the performance of the algorithm when only a small number of members of the protein family are available for training. Random subsets of the original GPCR training data set of size 400, 200, 100, and 50 were taken. These sequences were combined with 1000 random non-GPCR proteins as described above. The algorithm was retrained on the smaller data sets and tested on the test data set. This procedure was repeated 10 times for each sub-sampling size. The average positive identification rate and false positive rate is shown in Table 4. Even with only 50 training sequences we still obtained 91\% identification and only 0.5\% false positive identification.

Table 4. Test of the QFC algorithm for performance when smaller numbers of sequences are available for training. The first column shows the number of GPCR sequences used for training (see text for details). The second column and third column show the positive identification rate and false positive rate, respectively, for an independent test data set

<table>
<thead>
<tr>
<th># Training sequences</th>
<th>Positive identification (%)</th>
<th>False positive (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>400</td>
<td>98.1 (0.589)</td>
<td>0</td>
</tr>
<tr>
<td>200</td>
<td>95.1 (1.474)</td>
<td>0.1</td>
</tr>
<tr>
<td>100</td>
<td>93.6 (1.865)</td>
<td>0.2</td>
</tr>
<tr>
<td>50</td>
<td>91.3 (2.512)</td>
<td>0.5</td>
</tr>
</tbody>
</table>

Application to the Drosophila genomic database

Since the QFC algorithm does not rely on specific sequence patterns, it is not subject to sampling bias related to sequence motifs and sequence level homologies. Therefore, it is particularly suited to discovering novel multi-transmembrane proteins. The motivation of our project was to apply the QFC algorithm to discovering olfactory receptor genes in Drosophila. There have been extensive efforts to identify odorant and pheromone receptors in a variety of insects using a wide range of strategies. Nevertheless, search strategies based on sequence similarities have all been unsuccessful despite the fact that putative odorant receptor genes are available from vertebrates and C.elegans (Skoufos et al., 1999). The QFC
algorithm was applied to screen genomic sequences and ESTs from an earlier dataset in the Berkeley Drosophila Genome Project (BDGP, http://www.fruitfly.org; version available at 6/98) in an attempt to identify novel olfactory receptor genes in Drosophila (Clyne et al., 1999).

The QFC algorithm identified several hundred candidates from the genomic database after extracting open reading frames (ORFs) longer than 300 bases (translated in all six frames). (An earlier version of the algorithm trained on a more restricted set of data was used at this stage of our project.) This list of candidates was further reduced by comparing the candidate sequences against the Drosophila codon usage bias tables (http://flybase.bio.indiana.edu; version 10). Known genes were also discarded and 34 final candidate genes were identified. Most of the candidate sequences displayed fewer than seven putative transmembrane domains. Therefore, genomic sequences surrounding the candidate sequences were examined for the presence of neighboring ORFs that could be spliced to the candidate ORF, thereby increasing the size of the putative protein. We used Drosophila 5′ and 3′ intron–exon consensus splice signals to help identify linked exons (Mount et al., 1992). This procedure yielded several genes that could encode seven-transmembrane proteins. Laboratory work using RT-PCR analysis to examine tissue-specificity of expression revealed two strong candidates for olfactory receptor genes (Figure 3). These genes were used with the BLAST (Altschul et al., 1990) program to search the database for similar sequences, and the search identified 16 candidate genes from ~15% of the Drosophila genome as described in Clyne et al. (1999). As anticipated, the newly identified genes represent a completely new class of GPCRs that do not show significant sequence similarity to previously identified GPCRs. The closest sequence identified from a BLAST search to the sequences shown in Figure 3 is a human tyrosine protein kinase that shows 19% maximal match region within a 136 amino acid subsequence (expected value of 5.5 Karlin and Altshul, 1990).

Discussion

We have shown that our new QFC algorithm is fast, accurate—even for short subsequences of proteins, and useful in identifying an entirely new class of GPCRs. It also discriminate between GPCRs and other transmembrane proteins—as well with re-training it was reasonably successful at identifying other multi-transmembrane families such as ion channels. The main strength of the algorithm is that it is a classification scheme based on the feature space of protein structure. Most previous classification and identification algorithms have been based on sequence identity. Conversely, the main goals of previous structure-prediction algorithms have been to analyze structural features of a single molecule. The novelty of the QFC algorithm is that it is based on structural features shared by the entire class of multi-transmembrane proteins, specifically parameterized for GPCRs, and the characterization is used as an automatic classification and identification tool. This utility for ‘jumping’ protein families was recently demonstrated by its use in isolating Drosophila taste receptors (Clyne et al., 2000).

There are two key reasons why classification by structural features can be more useful than classification by sequence identity. In general, for many proteins their structural properties are likely to be more functionally important than the particular identity of the amino acids that comprise the protein. In this case, we would expect the structural properties to be evolutionarily conserved even when sequences diverge to an extreme state. For this reason structural information is being increasingly applied to establishing evolutionary relationships between proteins (e.g. Thorne et al., 1996). The second reason is that sometimes a functional classification of proteins is more useful than a genealogical or evolutionary classification of the protein. These two kinds of classification are often related but not necessarily so (thus the distinction between ‘homology’ and ‘analogy’). Again, by the ‘structure = function’ assumption, we are more likely to obtain a functional classification when we base it on structural features as compared to sequence identity.

Although the QFC algorithm has a slightly higher false positive rate than the other four existing methods, false positive rates can be reduced with secondary filters such as the codon bias filter we used for Drosophila sequences. Other filters might be developed, perhaps specific to particular multi-transmembrane proteins such as ion-channels. Such filters might include detection of unusual repeats, particular amino acid composition of the transmembrane domain, as well as the presence of other molecular signals like splice signals. (Currently, projects are underway in our labs to provide a computer analysis platform that incorporates all of these ideas; a preliminary version of the program used for the Drosophila database search is available from JK.) In general, a reasonable strategy would be to use our algorithm for a preliminary search and then use secondary searches that are more sequence-specific.

Overall, we believe our structural-statistical approach will be useful for rapid preliminary screening of large databases. We noted that the quasi-periodic properties of multi-transmembrane proteins were used as our main identifying features. Obviously, similar feature extraction can be carried out using more model-theoretic approaches such as using Hidden Markov models (e.g. Baldi and Chauvin, 1994). However, when there is sufficient biological information, e.g. the well-established quasi-periodicity of GPCRs, it seems best to directly
incorporate it into the structure of the statistical estimator rather than over-parameterizing with a complex model. Also, the main advantage of using a feature space + classification approach occurs when the feature space is constructed in such a manner that extrapolation and interpolation in the space makes biological sense. That is, suppose we know that two points A and B in our feature space are both GPCRs. Then it would be useful if all points in the line between A and B also corresponded to GPCRs. Since the quasi-periodic structure is an important functional component of GPCRs, it seems that using this structure to construct the feature space makes the most biological sense. This seems to be born out by our identification of a new class of odorant receptors (Clyne et al., 1999) and taste receptors (Clyne et al., 2000). Classification based methods hold great promise in terms of computational efficiency and statistical power. As pointed out by Jaakkola et al. (1999) the key lies in the construction of the proper feature space. For GPCRs and other multi-transmembrane proteins, measurements related to their quasi-periodic structure seem the most promising characterization.

Acknowledgements

We thank the members of the Carlson lab and Kim lab for discussions and advice. This work was supported by an Intel Technology for Education 2000 grant and a Merck Technology for Education 2000 grant and a Merck SCREENING utilizing a low order neural network filter. BioTechniques, 21, 1110–1114.


Kulp,D., Haussler,D., Reese,M.G. and Eeckman,F.H. (1997) In-

References


Kulp,D., Haussler,D., Reese,M.G. and Eeckman,F.H. (1997) In-

J.Kim et al.
Discriminant function detection of membrane proteins


