Automated generation and refinement of protein signatures: case study with G-protein coupled receptors

Michael I. Sadowski and John Howard Parish*

School of Biochemistry and Molecular Biology, The University of Leeds, Leeds LS2 9JT, UK

Received on June 19, 2002; revised on October 28, 2002; accepted on November 27, 2002

ABSTRACT

Motivation: Previous work had established that it was possible to derive sparse signatures (essentially sequence-length motifs) by examining points of contact between residues in proteins of known three-dimensional (3D) structure. Many interesting protein families have very little tertiary structural information. Methods for deriving signatures using only primary and secondary-structural information were therefore developed.

Results: Two methods for deriving protein signatures using protein sequence information and predicted secondary-structures are described. One method is based on a scoring approach, the other on the Genetic Algorithm (GA). The effectiveness of the method was tested on the superfamily of GPCRs and compared with the established hidden Markov model (HMM) method. The signature method is shown to perform well, detecting 68% of superfamily members before the first false positive sequence and detecting several distant relationships.

Contact: howard@bmb.leeds.ac.uk

Supplementary information: Software developed for this project and further data are available online (http://bbsrc-bioinf.leeds.ac.uk/BIOINF/jhp/)

INTRODUCTION

Previously, we described a method for aligning a sparse 'signature' descriptive of a family of proteins to a database of protein sequences (Daniel et al., 1999). We later demonstrated the diagnostic power of this method for a variety of families using signatures created with residue contact information derived from a three-dimensional (3D) structural alignment of representative family members (Ison et al., 2000). The difference between this approach and conventional motifs such as PROSITE (Hofmann et al., 1999) and PRINTS (Attwood et al., 2000) is that we specifically used structural information and, due to the 'elastic alignment' algorithm (Daniel et al., 1999) sparse signatures could represent sequence-length motifs. However, the earlier papers did not describe automated methods for generating such signatures. Here we describe a method for the extraction of signatures from sequences representing a family for which the structural information is limited, using the G-protein coupled receptors as an example of such a family. Specifically we review possible methods for the automated generation of signatures.

Overall the mean pairwise sequence-similarity of the GPCRs is approximately 15% (Horn et al., 2000a). There has been uncertainty about the sub-classification of GPCRs and their relationship to other groups receptors (such as bride of sevenless, BOSS, and the frizzled family). Josefsson (1999) examined the effects of seeding PSI-BLAST searches with the interior portions (sequences ranging from the start of the first putative trans-membrane domain to the end of the seventh) of some of these sequences and creating trees from significant results of these searches. He concluded that a relationship between the sequences of class 1 (rhodopsin-like), 2 (secretin-like), 5 (STE2 yeast pheromone receptors), 6 (slime-mold cAMP receptors) and the frizzled family could be detected in this way. BOSS and family 3 (metabotropic glutamate/calcitonin) receptors and the STE2 yeast pheromone receptors appeared as separate groups.

The classification of GPCR sequences in SWISSPROT (Bairoch and Apweiler, 2000) follows that of the GPCRdb (Horn et al., 1998), classifying sequences into five families, including the STE2 and STE3 yeast pheromone sequences in the same group. The distribution of known sequences within these families is heavily biased towards the rhodopsin-like family (family 1). This family accounts for over 90% of the GPCR sequences found in SWISS-PROT.

Despite their enormous diversity, the GPCRs are known to possess the same overall architecture (Sakmar et al.,...
2002); hydrophobicity analysis predicts that they all possess seven \( \alpha \)-helical trans-membrane regions, with an extracellular N-terminus and an intracellular C-terminus (Dohlman et al., 1991; Baldwin, 1994). These seven putative trans-membrane domains are the basis of the majority of the sequence-based analyses published about these receptors numbers of sequences (Donnelly et al., 1994). In contrast, the loop regions may vary greatly in size and composition (Horn et al., 1991; Baldwin, 1994). There is also a growing body of evidence that particular family 1 GPCRs can couple to a variety of G-proteins, sometimes in response to a variety of different ligands (Gudermann et al., 1997; Pauwels, 2000; Horn et al., 2000b). Recently, the first X-ray structure for a GPCR has been reported (Palczewski et al., 2000).

A great deal of research has been devoted to the elucidation of the functional and structural roles of various regions of representative GPCR sequences. Site-directed mutagenesis (Strader et al., 1994; Elling et al., 1997; Le Gouill et al., 1999; Scholl and Wells, 2000) and truncated receptor studies (Schoneberg et al., 1995; Ridge et al., 1996; Barbier et al., 1998; Scarselli et al., 2000) have elucidated the broad functional importance of specific residues and particular receptor regions in interactions with other signal transduction proteins.

**SYSTEM AND METHODS**

The aim was to create an automatic method for extracting sequence-length motifs for families of membrane proteins. The basic process is summarised in Figure 1.

**Derivation of Training and Target Sets**

The subset of entries in the SWISS-PROT release 38.0 containing the keyword ‘G-protein coupled receptor’ and not marked as fragments was extracted by a text parsing method. This set amounted to 957 sequences. This set of sequences was then reduced to 84 sequences by removal of all sequences more than 50% similar to one another from the set. Further reduction to 30% sequence identity was then performed, reducing the set to seven sequences, ‘S1’ (Table 1). A modified version of this set, with one of the seven sequences (ACM1_DROME) removed was used in the selection experiments, because this sequence contains a very large third intracellular loop (ca. 400 residues). This loop introduces the requirement for very large gaps in signatures to be permitted, increasing both the space and time requirements of the program considerably. The target set for identification consisted of 1276 GPCR sequences from SWISS-PROT 39.0, created by extracting all entries with the keyword ‘G-protein coupled receptor’.

**Creation of Multiple Alignments**

Four methods of multiple alignment were used. In the first (‘default’), a multiple alignment of the sequences of the training set taken wholly was created using ClustalW (Thompson et al., 1994) with default parameters. Trans-membrane sequences were predicted using TMAP (Persson and Argos, 1994). This alignment uses the matrices from the Gonnet series and will hereafter be referred to as the default (or standard) ClustalW alignment.

In the second method (‘P2’), sequences were split into regions of equivalent secondary structure by an automatic method using the annotations to SWISS-PROT to assign portions of the sequence as equivalent ‘features’ prior to alignment with CLUSTALW in the default way. In the present case this resulted in fifteen separate alignments of
putative loop and trans-membrane regions. These separate alignments were then recombined to create a full multiple alignment of the training set. Transmembrane regions were then annotated using TMAP in order to retain some equivalence between methods.

In the third and fourth methods (‘P3’, ‘P4’), the sequences were aligned as above but using a trans-membrane-specific substitution matrix in the case of aligning trans-membrane regions. The matrices used were the PHAT 8588 matrix (Ng et al., 2000) and the SLIM 160 matrix (Muller et al., 2001) respectively.

**SIGNATURE Alignment Software**

The signature method (Daniel et al., 1999) is related to other methods of sequence identification, but differs in detail from all of them. A signature is effectively a motif for an entire sequence of a certain type. This motif consists of potential residue positions separated by flexible gaps which are allowed to adopt certain specified lengths without penalty. Further extension or compression outside these limits is penalised with an affine gap penalty as in sequence-similarity searching.

The software of Daniel et al. (1999) has been modified to simplify the input parameters and to remove a small design error in the original code. The revised version of SIGNATURE and the software used for generating such files of sequences are all available as source code from our URL (Supplementary Information).

**Statistical Significance of Signature Scoring**

To assess the performance of the signatures, the score corresponding to a 1% significance level for a given signature with a given set of parameters was calculated by a simulation method. Each signature was aligned to a database of 50 000 randomly-generated protein sequences with randomly-generated lengths (within the limits used for searching the SwissProt database). The sequences were generated by a zero-order Markov process using the residue frequencies derived from the SwissProt 22 database for the creation of the JTT matrices (Jones et al., 1994). The scoring level at which 99% of the database members were found was then derived numerically and used as a threshold for the evaluation of database searches.

Sequences exceeding this score were then counted as ‘hits’ and sensitivity and selectivity of the signatures were calculated according to the following formulae:

\[
\text{Sensitivity} = \frac{TP}{TP + FN}
\]

\[
\text{Selectivity} = \frac{TP}{TP + FP}
\]

Where \( TP = \text{true positives, } FN = \text{false negatives and } FP = \text{false positives.}\)

**Construction of Hidden Markov Models**

In order to compare the diagnostic power with a very successful measure and to determine the efficacy of the constrained (‘P2’, ‘P3’ and ‘P4’) alignment procedures hidden Markov models (HMMs) were created using the HMMer program (Version 2.1.1; available from http://hmmer.wustl.edu) (Eddy, 1998) from each of the four alignments created for use in the selection process. The models were calibrated using the calibration software provided with the HMMer release, and used to scan the SWISS-PROT database. The HMMs derived were then compared with the best signature found for each alignment.

**ALGORITHMS**

Two algorithms were developed for the purposes of key residue selection. One method re-scores multiple alignments using substitution matrices. The other is an adaptation of the Genetic Algorithm for the purpose of motif selection.

**Selection of Key Residues by Substitution Matrix**

Aligned positions were scored using a substitution matrix in the following way: each residue in turn was scored against all the other residues (excluding itself) in the aligned position and the total score was divided by the number of comparisons made, being equal to \( n(n-1)/2 \) (where \( n \) is the number of aligned sequences). The score was then compared to a pre-specified threshold and the residue marked as ‘key’ if the threshold was exceeded.

Selections were calibrated on the basis of their stringency, defined as the number of positions selected by each matrix using a given threshold, such that selections of approximately equal stringency (due to entanglement between positions under a given scoring system it was not always possible to select exactly equal numbers of positions by the various methods) could be compared to one-another.

The following substitution matrices were compared using selections from the ‘CL’, ‘P2’ and ‘P3’ alignments: the BLOSUM 40 matrix (Henikoff and Henikoff, 1992); the Risler matrix (Risler et al., 1988); the JTT 50 matrix (Jones et al., 1994) the PHAT 8589 matrix (Ng et al., 2000) and the SLIM 160 matrix (Muller et al., 2001).

Unconstrained selections of approximately 48, 72, 96, 120, 144, 168, 192 and 216 positions were made; from the TM regions, selections of approximately 11, 22, 33, 44, 55, 66, 77, 88 and 99 positions were made with each matrix. The ‘P4’ alignment was also tested using unconstrained selection and the JTT 50, BLOSUM 40 and Risler matrices for

G-protein coupled receptors
selection. In addition, selections were made from loop regions only and the trans-membrane regions plus the third intracelluar loop and C-terminus.

The effect of the use of different matrices for aligning to the signatures to the database was also examined. The matrices used for selection as well as the BLOSUM 62 and identity matrices were all compared with the signatures derived from the ‘P3’ alignment.

Software written for selection of the key residues (‘SKEY’) and for alignment of regions (‘SSALIGN’) is available from our URL (Supplementary Information).

Selection by Genetic Algorithm

Key residue selection was also performed using a Genetic Algorithm (GA). Randomisation relied on the ran2 function of Press et al. (1993). The software, ‘OPTIMISE4’, is available as source code from our URL (Supplementary Information). The encoding used in the genetic algorithm is illustrated in Figure 2.

The population of solutions acted on by the GA consisted of strings of characters from the set \{1,0\}. Strings of length equivalent to the number of non-gap alignment columns were generated at random. The encoding was then interpreted as ‘1’ meaning that the position should be selected, ‘0’ that it should not. An initial population of these strings was then generated at random.

To calculate the fitness of a population member, the member was first translated into a signature and used to scan the SwissProt 39.0 database. The fitness was then calculated as the percentage of true positives occurring before a pre-specified rank in the list of top-scoring sequences. In all cases this rank was the number of true positive sequences in the database plus one, 1277 in the case of the G-protein coupled receptors.

Reproduction consisted of sexual and asexual components. Asexual reproduction consisted of insertion of exact duplicates into the new population. Sexual reproduction used a simple one-point crossover strategy. In both cases the probability of reproduction was controlled by fitness. Mutation (by an XOR 1 operation) was performed over all loci in the population such that each had a 0.5% probability of mutating.

The GA was run twice with a population size of 100 for 100 generations. In one case selection was permitted from all regions of sequence, in the second only positions from regions predicted by TMAP to be membrane-spanning were available for selection.

It was necessary in both cases to exclude columns containing gaps from the selection process as the current version of the signature software requires an equal number of amino-acids to be present at each position. Gap opening and extension penalties were both set at 11 for intra-SSE positions and 10 for inter-SSE positions in all cases.

IMPLEMENTATION

All of the software described was implemented in C or C++ on LINUX and BSD UNIX systems. Inputs to the selection processes consist of multiple sequence alignments in MSF format and secondary structural predictions. Output in each case is a ‘key residue’ file.

RESULTS

Derivation of signatures

333 signatures were derived by using the combinations of matrices and constraints described above. Parts of a signature file (supplem.pdf) as well as complete files of this type are available from our URL (Supplementary Information).

Effect of Alignment Procedure on Detection

Effects of alignment on detection were analysed by coverage versus error plots (Supplementary information: supplem.pdf). In the cases both of signatures and of HMMs it was found that the second alignment procedure resulted in an improved level of detection. However, the use of the transmembrane-specific PHAT matrix to align transmembrane regions separately resulted in an increase in both sensitivity and selectivity of the signature method but greatly decreased sensitivity of the HMM method. The use of the SLIM 160 matrix resulted in an increase in discriminatory power both for signature and HMMer and proved the best alignment procedure overall. The default ClustalW alignment proved extremely poor as a basis for the derivation of signatures, but produced a reasonable HMM.

Selection with substitution matrices

Selection using the JTT 50 matrix was the most consistently well-performing method, particularly on the basis of signature performance at more stringent selection levels. The Risler matrix proved a particularly poor basis for selection of key residues, producing signatures which were both less sensitive and less selective than those derived using other matrices, and which decreased in overall performance with increasing selection stringency. Overall it was possible to select reasonable signatures with any of the other matrices tested. In general the peak of performance was observed when approximately 90 positions from the multiple alignment were selected.

EFFECT OF CONSTRaining SELECTION ON DETECTION

Selection from transmembrane regions was better than either of the other procedures but not quite as powerful as unconstrained selection. The third set of regions (TM + two ‘loop regions’) was very close in performance to the unconstrained set in the majority of cases, sometimes
Fig. 2. Selection by Genetic Algorithm. In the upper panel, the alignment is shown and gap positions are in bold and these are assigned values in a binary string referred to as a population member (Pop. Mem.). Initially a random population of these strings is generated and each is translated into a motif which is aligned to sequences in the database to provide a fitness value for the string. Recombination and mutation generate a new population according to fitness values. The procedure is repeated until a specified time limit or fitness level is reached. The alignment shown is a section of the P3 alignment for the S1 training set.

performing slightly better. Selection from putative loop regions only performed very poorly except in a few cases where selection was very stringent. In these cases it seemed to be where residues from the 7th transmembrane region (which was not detected by TMAP in any case) and the cytoplasmic ‘H8’ region were being selected, as these regions contained the most conserved positions.

Effect of Scoring Matrix on Alignment

It was found that the PHAT, JTT 50 and BLOSUM matrices performed similarly with respect to their sensitivity scores, BLOSUM 40 and PHAT being the joint highest scoring, with the BLOSUM 62 and JTT 50 matrices slightly lower. The identity and Risler matrices both yielded much lower sensitivity scores than the other four (Supplementary information: supplem.pdf). The SLIM matrix performed poorly with respect to both characteristics (data not shown).

Analysis of GA Selection Results

The results of a typical GA run are summarised in Figure 3; positions in the multiple alignment are annotated according to their tendency to appear in the final population after 100 generations of selection. Only the first six trans-membrane helices are marked in the figure because helix 7 was not reliably predicted by the TMAP software. The illustrated run was performed for selection from the entire sequence, and it can be seen from the figure that the pattern of strongly conserved alignment positions favours positions predicted to lie in the first three helices. Interestingly the [DE]RY motif which occurs at the cytoplasmic end of helix III in rhodopsin family GPCRs is not strongly conserved in the case of selection from transmembrane regions only, although a residue which occurs immediately afterwards exhibits a strong presence in the population. In the case of unconstrained selection the [DE]R part of this region is conserved, however (Supplementary information: supplem.pdf). The signature derived by selection from all regions of the sequence was 84% sensitive and 78% selective. This puts it roughly on a par with the fitness of the highest-scoring signatures selected from the same alignment, although in this case sensitivity has been biased for over selectivity by nature of the fitness function used. Derivation from TM regions only resulted in a signature that was similarly TM but much less selective (54%).

Detection of GPCR Family Members

The detection of members of the five families of the GPCRs by the best signature derived so far is summarised in Table 2.

The pattern of family members detected by this signature is representative of more selective signatures (this example has a sensitivity of 78% and selectivity of 91%, detecting 994 GPCR sequences along with 97 false positive sequences.). A few representatives of the *D.melanogaster* olfactory receptor family sequences were also present in most cases. In some cases representatives of family 2 or (more rarely) family 3 also appeared in the
subset of detected sequences, although in all of the cases that were examined they appeared close to the cutoff point determined as the threshold for detection.

Comparison with PROSITE motifs

The alignment of the best signature to the GPCR sequences was compared to that of the PROSITE motifs for families 1, 2, and 3. The comparison is shown in Table 3. Six of the eleven positions of the family 1 motif (PS00237) were matched in better than 94% of the signature alignments to those sequences. Although families 2 and 3 were not well detected by this method, some positions in their motifs were also detected by signature in more than 60% of family members. The second family 2 motif (PS00650) was well detected, the family 3 motifs less so. As expected the first family 3 motif (PS00981) was not detected by the signature, as it is located within the extracellular N-terminal domain of these proteins.

CONCLUSIONS

Our objectives were to establish whether the signature concept was applicable to cases where the structural information was limited to predicted secondary structure elements and whether there are methods available to automate the generation and refinement of such signatures. We conclude that both of these are the case.

The matrix-based selection method presented here suffers from the ‘curse of dimensionality’ with regard to the large number of parameters required. Use of the genetic algorithm selection method is therefore preferable, although the present method requires large amounts of computer time (roughly three months for 100 generations of 100 population members) owing to the large size of the database.
The overall performance of the signatures derived in this study suggests a trade-off between sensitivity and selectivity (or specificity) of classification. The results of this study suggest a limit of 80% coverage at a 1% error rate for this method. The results of HMM detection using the same alignment suggest a ceiling for this set of 85% coverage with 1% error.

Griffiths-Jones and Bateman (2002) present work which demonstrates that use of structural alignments as a basis for creation of HMMs does not offer an improvement in levels of homologue detection over purely sequence-based multiple alignment methods. The apparent contradiction with our own results can be resolved by considering that over such a large evolutionary distance (possibly containing examples of convergence, although the lack of significant sequence-similarity between, e.g. families 1 and 3 does not necessarily dictate that this is the case) it may not be the accuracy of specific parts of the alignment so much as its accuracy over broad regions which is important. The improved prediction of transmembrane helices by TMAP suggests that this alignment procedure improves regional composition as expected.

How many sequences are required for the characterisation of a sequence family is still an open question and one which is very difficult to rigorously assess (even assessing all possible sets of any six sequences requires \(4 \times 10^{18}\) combinations). The ability of signatures to identify a good proportion of this family at quite some distance using such little information seems to imply that redundancy exists in proteins both at the sequence and the family level. In other words, there is something to be gained from ignoring the information contained both in a large proportion of the sequences in the family and within a large proportion (somewhere up to 80%) of the sequence.

ACKNOWLEDGEMENTS

MIS is supported by the Medical Research Council. We thank Dr D.R. Westhead and Mr M.J. Blades for helpful discussions. The authors are very grateful Richard Ewan for his contribution to the preliminary stages of this study.

REFERENCES


M.I. Sadowski and J.H. Parish


