



Prediction of the coupling specificity of G protein coupled receptors to their G proteins

Steffen Möller¹, Jaak Vilo¹ and Michael D.R. Croning^{1,2}

¹EMBL-EBI, European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SD, UK and ²School of Biological Sciences, The University of Manchester, Oxford Road, Manchester, M13 9PT, UK

Received on February 5, 2001; revised and accepted on April 1, 2001

ABSTRACT

G protein coupled receptors (GPCRs) are found in great numbers in most eukaryotic genomes. They are responsible for sensing a staggering variety of structurally diverse ligands, with their activation resulting in the initiation of a variety of cellular signalling cascades. The physiological response that is observed following receptor activation is governed by the guanine nucleotide-binding proteins (G proteins) to which a particular receptor chooses to couple. Previous investigations have demonstrated that the specificity of the receptor-G protein interaction is governed by the intracellular domains of the receptor. Despite many studies it has proven very difficult to predict *de novo*, from the receptor sequence alone, the G proteins to which a GPCR is most likely to couple. We have used a data-mining approach, combining pattern discovery with membrane topology prediction, to find patterns of amino acid residues in the intracellular domains of GPCR sequences that are specific for coupling to a particular functional class of G proteins. A prediction system was then built, being based upon these discovered patterns. We can report this approach was successful in the prediction of G protein coupling specificity of unknown sequences. Such predictions should be of great use in providing *in silico* characterisation of newly cloned receptor sequences and for improving the annotation of GPCRs stored in protein sequence databases. Available at: <http://www.ebi.ac.uk/~croning/coupling.html>

Contact: croning@ebi.ac.uk

INTRODUCTION

GPCRs are the biggest single class of receptors in biology, playing key roles in a remarkably wide range of physiological and pathophysiological conditions. The actions of a large and structurally diverse range of hormones, neurotransmitters, tastants, odourants, photons, and peptidases, are initiated by their binding to GPCRs located on the cell surface (Bockaert and Pin 1999). Such binding activates the receptor, causing helical rearrangements within the

receptor, which (by way of unmasking binding sites) transmits the activation signal to a guanine nucleotide-binding protein (G protein) located on the cytoplasmic surface of the membrane, closely apposed to the receptor (Schoneberg, Schultz et al. 1999; Gether 2000).

Activation of the heterotrimeric G protein (consisting of α , β , and γ subunits) promotes exchange of the guanosine diphosphate (GDP), bound to the α subunit, for guanosine triphosphate (GTP). This allows the dissociation of the α subunit (with GTP bound) from both the receptor and $\beta\gamma$ complex. The separate moieties can then modulate several cell signalling pathways, and the activities of certain ion channels. Termination of the response occurs as a result of the intrinsic catalytic activity of the α subunit, which hydrolyses the bound GTP to GDP. Subsequently the α -GDP then re-associates with the $\beta\gamma$ complex to form the inactive heterotrimer. Amongst the biochemical responses that have been observed following receptor activation (LeVine 1999) are both stimulation and inhibition of adenylate cyclase activity. The G_s class, and the $G_{i/o}$ class of G proteins, respectively, mediate these opposing effects. The $G_{q/11}$ family activate phospholipase C enzymes, resulting in phosphatidylinositol hydrolysis. Together these three families constitute the major functional classes of G proteins, and studies have revealed this specificity is determined by the particular subtype of the α subunit, making up the G protein (Simon, Strathmann et al. 1991; Bourne 1997).

Characteristically each GPCR subtype appears to only couple to a subset of the G proteins that may be found in a particular cell. Elucidation of the mechanism(s) underlying this coupling specificity has been a central theme in GPCR research over the last 15 years. Biochemical studies, especially those that involve the creation of chimeric receptors, have been used in order to locate domains within receptor sequences that may define their specificity of G protein coupling. Other strategies that have been employed are the use of synthetic peptides, which are designed to mimic or inhibit the normal receptor-G protein interactions, and the neutralisation of

specific G proteins with antibodies. Together this large number of studies have revealed that the selectivity of G protein recognition (and hence coupling) is determined by multiple intracellular receptor regions. The most important regions appear to be the second intracellular loop, and the start and end of the third intracellular loop, which are close to the cytoplasmic surface of the membrane (Wess 1998).

However, the coupling specificity has yet to be experimentally determined for many hundreds of mammalian GPCRs, including many peptide receptors (Liu and Wess 1996). This knowledge is important for two main reasons, firstly, to understand the physiological mechanisms underlying the response mediated by activation of a given GPCR, and secondly, in order to choose appropriate cell lines for the heterologous expression of newly-cloned GPCRs. This is crucial for the study of the increasing catalogue of GPCRs that have been cloned but for which the endogenous agonist is unknown, the so-called orphan receptors (Wilson, Bergsma et al. 1998). For the ligand-identification strategy that is applied to these orphans depends upon functional coupling of the receptor to a G protein, so that a downstream change (such as a change in second messenger concentration), can be observed with a suitable assay. One then passes appropriate tissue extracts (or libraries of chemical compounds) over the cells, hoping to observe a response. Of course for such an identification method to succeed G-proteins must be present in the chosen cell, to which the receptor is willing to couple. In an effort to improve this likelihood, a number of transgenic systems have been developed, by introducing native or engineered G protein α subunits that are promiscuous in their coupling to receptors (Wess 1998). Clearly the development of an accurate method for the prediction of coupling specificity of a receptor to G protein(s) would be of great utility in guiding experimental investigations for the characterisation of GPCRs. Since no receptor sequence motifs have been reported that unambiguously determine coupling specificity across GPCR families and subtypes, we thought it unlikely with respect to our simple human inspection of GPCR sequences, that we would be able to develop a successful prediction system. Instead we decided to use the prior knowledge that the sequences motifs we were seeking would likely be located in the intracellular domains of the receptor, together with a protein pattern discovery algorithm to hunt for commonly occurring patterns in these intracellular loops and C-termini. Following the identification of a large number of patterns, we then evaluated their usefulness in prediction, based upon a set of approximately 100 paralogous receptor sequences for which the G protein coupling specificity had previously been experimentally determined. We can report that using such a bioinformatics approach (combining

membrane topology prediction with pattern discovery) one can discover combinations of patterns that are indeed characteristic for the coupling of receptors to G proteins.

METHODS

Our strategy to predict the coupling specificity of GPCRs for their G proteins was to attempt to find patterns of amino acid residues in their sequences that appeared to be specific to a particular class of G protein. In order to do this we required a set of GPCR sequences for which the coupling specificity has been reported, and a pattern discovery algorithm. We retrieved 103 diverse receptor sequences from SWISS-PROT and TrEMBL for which an apparently non-promiscuous coupling had been determined and was summarised in the TiPS Nomenclature Supplement (Bairoch and Apweiler 2000; TiPS 2000). These were grouped into the three functional classes $G_{i/o}$, G_s and $G_{q/11}$. To constrain the search for patterns to the putative intracellular domains of the sequences, we required an accurate membrane topology prediction. In a previous study we demonstrated that TMHMM is the best currently available topology prediction method, for determining the membrane spanning regions (MSRs) of GPCR sequences (Möller, Croning et al. 2001). Here, we have used a modified version of this program, called 7TMHMM, which was designed specifically to predict the MSRs of GPCRs. The model employed assumes exactly 7 MSRs, with extracellular and intracellular, N- and C-termini, respectively (Möller, Croning et al. 2001).

Generation and evaluation of patterns

The pattern discovery was carried out using the program SPEXS (Vilo 1998). This performs an exhaustive search within the input sequences with regular expressions as the predefined pattern language. Amino acids were grouped by property as described in (Livingstone and Barton 1993). This produced a large number of patterns (~4000) which we then evaluated for their usefulness. The most discriminative patterns are those that occur in large number of sequences of one receptor-G protein coupling group and infrequently in the others. The specificity of all the patterns occurring in each group of receptor sequences were determined.

For every pattern we calculated the likelihood of its appearance in its respective receptor-G protein coupling group, normalised by its occurrence in all the sequences contained in the three functional classes of G protein coupling. The pattern score was calculated as the inverse of the probability, adapting an earlier method used to estimate the significance of patterns found in DNA sequences (Brazma, Jonassen et al. 1998; Vilo, Brazma et al. 2000). Thus, the smaller the probability, the higher the pattern score.

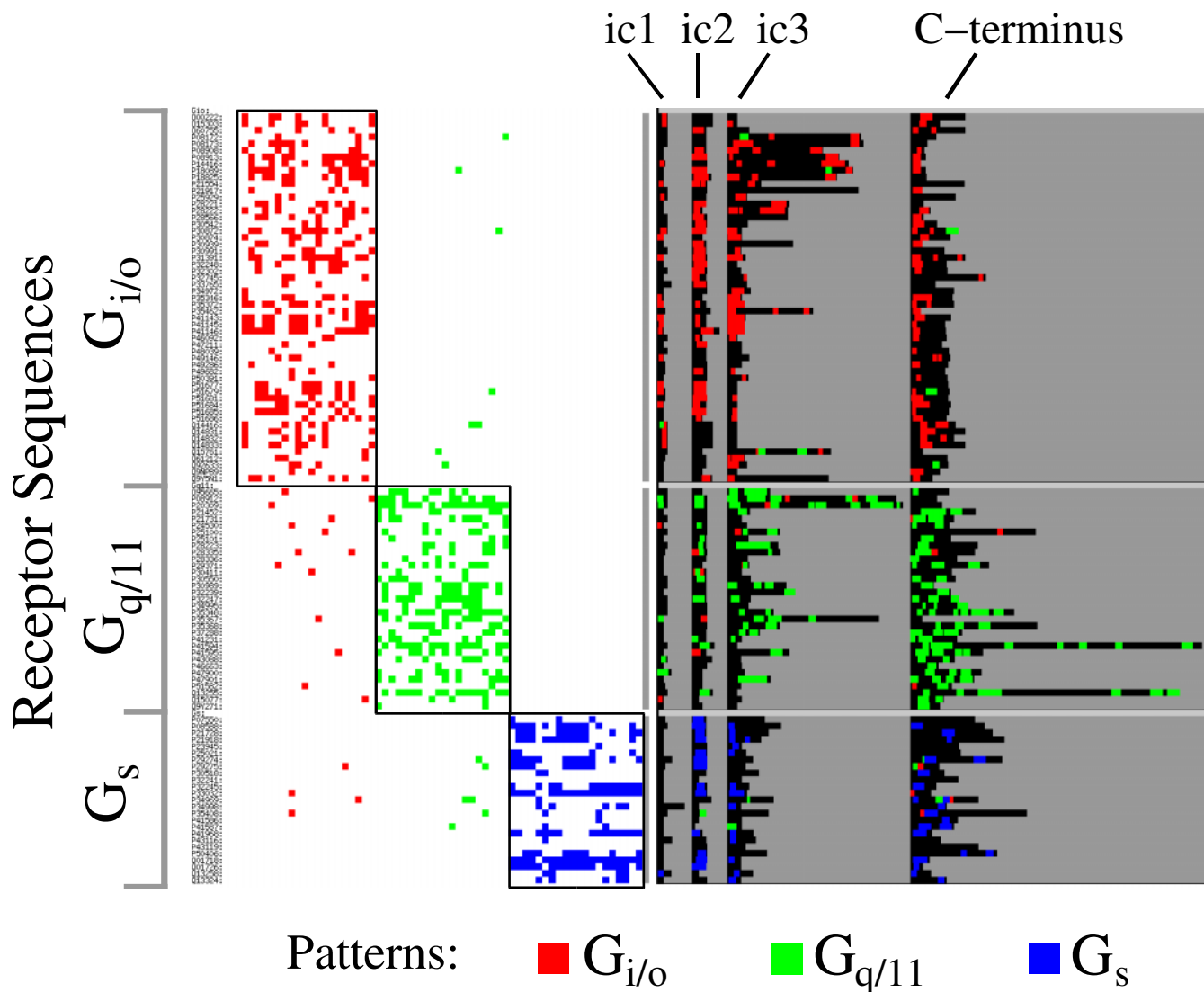


Fig. 1. Visualisation of pattern matches upon protein sequences within the training set. Each line represents an individual receptor protein, and these are grouped by their coupling specificity. The left-hand panel shows protein sequence accession numbers, and pattern matches that are coloured according to the receptor-G protein coupling group from which the patterns were discovered. The blocked regions reveal the majority of discovered patterns have the desired specificity. The right-hand panel shows the intracellular loop regions (ic1-ic3) and the C-termini of the receptors depicted as black bars that are proportional to the sequence lengths of these domains. Positions of pattern matches are highlighted upon them.

We hypothesised that we might improve the classification of the receptor-G protein coupling groups (and thus subsequent prediction of the coupling for a novel sequence) if we considered the specificity of combinations of patterns, rather than just single patterns. This is similar to the concept of collections of motifs (called fingerprints) that are found in the secondary protein database PRINTS (Attwood, Croning et al. 2000), or the analysis of the regulatory regions in DNA (Scherf, Klingenhoff et al. 2000). Derived pairs and triplets of patterns that are specific for

the binding to G proteins are used in conjunction to act as a classifier.

If all the patterns making up a particular combination were found in a sequence, the combination was said to match as a whole. For each sequence submitted to the classifier we report the total number of combinations found, and if 30% or more of the matches happened to belong to a specific receptor-G protein coupling group, then this coupling was assumed to be a putative prediction. This potentially allows one to predict promiscuous receptor-G

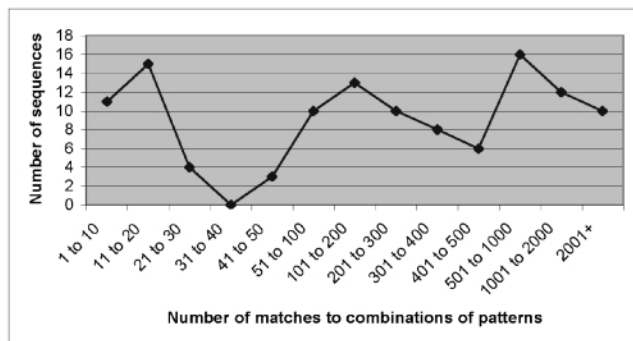


Fig. 2. Distribution of the number of combinations of pattern matches within the training set. The graph summarises how many of the sequences are found to have a particular total number of matches.

protein coupling. In order to test the resultant classifier, we predicted the G protein coupling specificity of 10 human GPCR subtypes not in our training set but for which the G protein coupling had been experimentally determined, or could be inferred from the biochemical responses seen following activation of the receptor.

RESULTS

Lists of patterns

Table 1–3 shows the 40 best patterns found for each receptor-G protein coupling group, together with the number of times they match in each of the three training set groups, and their calculated sensitivity and specificity. A visual inspection of the patterns confirmed our numerical analysis (as tabulated). The tool PATMATCH, which is part of the Expression Profiler package (Vilo et al., in preparation) was used for this purpose, and allowed us to visualise pattern matches upon the sequences, having grouped the latter by their G protein coupling specificity (see Figure 1). Most of the patterns were seen to match in just one of the three groups of receptor sequences, with few matches to the other two groups, demonstrating their usefulness and specificity. Additionally all of the GPCR sequences were matched by at least a few patterns. PATMATCH also allowed us to determine where the patterns matched on the intracellular domains of the receptor sequences, and from this to deduce whether match positions are conserved both within a particular receptor-G protein coupling group, and between the three groups.

Figure 2 shows that low numbers of matches are found for about 20% of the sequences in the training set. Particularly, match totals in the range 31–100 are of concern. This may have resulted from a rather limited number of patterns being found in these sequences. We did

not attempt to reduce redundancy in the selected patterns (Table 1–3), although this might have helped to reduce bias in the number of patterns available per sequence. Another possible confounding factor is an inaccuracy in the membrane topology prediction for particular receptor sequences, which would presumably constrain our pattern discovery to portions of the molecule that are unlikely to govern G protein coupling specificity. Our predictions are dependent upon the accuracy of the receptor-G protein coupling specificity information summarised in (TiPS 2000). We cannot exclude the possibility that some useful patterns might have been lost due to an incorrect or promiscuous coupling assignment.

Verification of classification on novel GPCR sequences

The classifier was applied to the sequences of 10 receptor subtypes not present in the training set, as shown in Table 4. Pairwise alignments revealed that these test sequences were in general 30–40% identical to their most similar paralogue in the training set. All 10 predictions appeared to be correct when we consulted the primary literature. With reference to Figure 2 we did not expect to trust predictions based on less than 50 matches. We were surprised that the predictions for both P41180 and P25105 were correct given that they resulted from a rather low number of matches.

In order to determine whether pattern discovery was strictly necessary for correct prediction, we also took the simpler approach of building a dendrogram from a multiple sequence alignment of the inner domains of the training set sequences. We did observe some propensity for receptors with the same coupling preference to be near each other in the tree, however, the delineation between the three groups of receptors was far from distinct.

DISCUSSION

We found a range of protein patterns within GPCR sequences which are apparently involved in determining their selectivity of binding to (and activating) the different functional classes of G proteins. We focused our pattern discovery upon the intracellular domains of GPCRs (previously reported to be involved in the receptor-G protein interaction) using a novel membrane topology prediction algorithm designed specifically for GPCRs. The resulting patterns were employed in a combinatorial manner to build a classifier, allowing one to predict the G protein coupling specificity of “unknown” receptor sequences that were not present in our training set.

The dependency of the classification on a prior determination of the analysed protein sequence as a GPCR implies a context-dependence for the usage of the patterns. This a-priori knowledge can be derived from protein domain

Table 1. List of generated patterns found to best represent a specific coupling mechanism. Column 1 shows the pattern as a regular expression, columns 2-4 show the number of matches to the different receptor-G protein coupling groups, columns 5 and 6 show sensitivity and specificity, respectively.

Pattern	G_{i0}	$G_{q/11}$	G_s	Sensitivity	Specificity	Best class
Total	55	33	25			
[ILV]...SG.{0,10}R	15	0	0	0.273	1	$G_{i/o}$
N..R.{1,4}R	15	0	0	0.273	1	$G_{i/o}$
Y.A.{1,8}A[ILV]	15	0	0	0.273	1	$G_{i/o}$
A[ILV].{2,5}RT	15	0	0	0.273	1	$G_{i/o}$
N..[RK]..R	17	1	0	0.309	0.9444	$G_{i/o}$
K.[RK].{0,10}K.[ILV]	17	1	0	0.309	0.9444	$G_{i/o}$
V...[RK]...R	17	1	0	0.309	0.9444	$G_{i/o}$
[RK]...[CM][RK]	23	1	2	0.418	0.8846	$G_{i/o}$
V[RK].{1,10}SG	16	1	0	0.291	0.9412	$G_{i/o}$
K.[RK].{1,4}L[RK]	16	1	0	0.291	0.9412	$G_{i/o}$
[FWY][ILV].V.{2,10}R	15	1	0	0.273	0.9375	$G_{i/o}$
Y.[RK].[RK].{0,9}T	15	1	0	0.273	0.9375	$G_{i/o}$
[ILV].A[AGS].{1,4}R	15	1	0	0.273	0.9375	$G_{i/o}$
FR...[RK].{0,3}L	15	1	0	0.273	0.9375	$G_{i/o}$
DRY.[AGS].{3,6}A	15	1	0	0.273	0.9375	$G_{i/o}$
F[RK]...K.{1,7}C	15	0	1	0.273	0.9375	$G_{i/o}$
A...[ILV].{1,8}RT	15	1	0	0.273	0.9375	$G_{i/o}$
[RK]...R.{0,9}EK	15	0	1	0.273	0.9375	$G_{i/o}$
[RK]R.{0,3}TR	15	1	0	0.273	0.9375	$G_{i/o}$
KA.{3,6}T	15	1	0	0.273	0.9375	$G_{i/o}$
DR.{4,11}H...[AGS]	15	1	0	0.273	0.9375	$G_{i/o}$
R...K.{0,8}T[AGS]	15	1	0	0.273	0.9375	$G_{i/o}$
[RK][FWY][ILV].{2,5}V	18	1	1	0.327	0.9000	$G_{i/o}$
N.{2,5}R.[FWY]	18	1	1	0.327	0.9000	$G_{i/o}$
Y.[AGS].{1,8}A[ILV]	18	2	0	0.327	0.9000	$G_{i/o}$
N..[RK].{1,4}R	23	3	1	0.418	0.8519	$G_{i/o}$
[ED].{0,3}N..[RK]	23	2	2	0.418	0.8519	$G_{i/o}$
Y.{2,5}I..[AGS]	23	0	4	0.418	0.8519	$G_{i/o}$
N..[RK].{1,11}R	30	6	2	0.545	0.7895	$G_{i/o}$
[RK].R.{2,12}K[RK]	20	4	0	0.364	0.8333	$G_{i/o}$
[ILV]...SG	20	1	2	0.364	0.8696	$G_{i/o}$
[AGS][RK]..[ED].{0,10}R	17	1	1	0.309	0.8947	$G_{i/o}$
[FWY].A.{1,9}A[ILV]	17	2	0	0.309	0.8947	$G_{i/o}$
R[FWY].[AGS][ILV].{0,7}A[ILV]	17	2	0	0.309	0.8947	$G_{i/o}$
[ILV].R...V	17	0	2	0.309	0.8947	$G_{i/o}$
[RK]Y.[AGS].{3,5}A	17	0	2	0.309	0.8947	$G_{i/o}$
[ILV]...SG.{0,8}E	17	0	2	0.309	0.8947	$G_{i/o}$
[FWY].[AGS][ILV]..A	17	1	1	0.309	0.8947	$G_{i/o}$
[RK]..[RK].{0,3}R[ILV]	32	8	2	0.582	0.7619	$G_{i/o}$
[ED]A.{0,3}E	19	3	0	0.345	0.8636	$G_{i/o}$

databases like PRINTS or PFAM, from the results of similarity searches, or can be read directly from the manual annotations present in databases such as SWISS-PROT. Additionally, the patterns should be employed in the context of the prediction of receptor sequence's membrane topology. With the increasing modularity of large-scale annotation efforts (Fleischmann, Möller et al. 1999; Möller, Leser et al. 1999; Birney 2001) such contextual information can now be technically incorporated into genome annotation. The present study thus represents an early example of a new breed of context-dependent protein domain annotation.

Clearly many aspects of the interaction between a receptor and its G protein(s) remain to be investigated. Our method of modelling whether a receptor is likely to be promiscuous in G protein coupling is straightforward. It would be worthwhile determining whether unique interaction motifs exist for promiscuous coupling in receptors that have been demonstrated to lack selectivity in their G protein interactions. We did not try to construct patterns for the exclusion of certain G protein couplings, i.e. a pattern to represent an exception to a rule. Our approach could eventually be improved by ignoring any pattern combination that does not span at least two inner

Table 2. List of generated patterns found to best represent a specific coupling mechanism. Column 1 shows the pattern as a regular expression, columns 2-4 show the number of matches to the different receptor-G protein coupling groups, columns 5 and 6 show sensitivity and specificity, respectively.

Pattern	G _{io}	G _{q/11}	G _s	Sensitivity	Specificity	Best class
Total	55	33	25			
T. [RK]. {0,10}S. .T	0	11	0	0.333	1	G _{q/11}
A. {3,6}V[ILV][RK]	0	11	0	0.333	1	G _{q/11}
P. [AGS]T. {0,10}S	0	10	0	0.303	1	G _{q/11}
[AGS][ILV][ILV][RK]. {2,10}S	0	10	0	0.303	1	G _{q/11}
S[FWY]. {1,11}Q[ILV]	0	10	0	0.303	1	G _{q/11}
[AGS]. {0,3}S. .T[ILV]	0	10	0	0.303	1	G _{q/11}
S. .L. {2,9}TL	0	10	0	0.303	1	G _{q/11}
[RK]F. . . .K	0	10	0	0.303	1	G _{q/11}
[AGS]. [ILV]. {0,10}K.F	0	10	0	0.303	1	G _{q/11}
[AGS]. S. [RK]. {0,10}F	1	13	0	0.394	0.9286	G _{q/11}
S. .L. {1,10}T[ILV]	1	12	0	0.364	0.9231	G _{q/11}
[RK]. T. {0,10}Q[AGS]	0	12	1	0.364	0.9231	G _{q/11}
[AGS]. . .L. {1,10}TL	1	12	0	0.364	0.9231	G _{q/11}
[AGS][ILV][ILV][RK]	0	12	1	0.364	0.9231	G _{q/11}
A. {0,10}V[ILV][RK]	1	14	1	0.424	0.8750	G _{q/11}
[AGS]. {0,3}V[ILV][RK]	1	14	1	0.424	0.8750	G _{q/11}
F. {0,10}Y. . . [RK]	0	14	2	0.424	0.8750	G _{q/11}
[CM]. [FWY]. {3,12}P	1	11	0	0.333	0.9167	G _{q/11}
S. [AGS]. {3,13}TL	1	11	0	0.333	0.9167	G _{q/11}
V[AGS]. {0,10}S. [AGS]. [ILV]	1	11	0	0.333	0.9167	G _{q/11}
Y. . . [RK]P. {2,10}A	0	11	0	0.333	1	G _{q/11}
[ILV].A. T	1	11	0	0.333	0.9167	G _{q/11}
S. .L. {1,11}Y	1	11	0	0.333	0.9167	G _{q/11}
A. {3,12}V[ILV][RK]	0	11	1	0.333	0.9167	G _{q/11}
[AGS]. {2,5}V[ILV][RK]	1	11	0	0.333	0.9167	G _{q/11}
[FWY]. {4,7}KP	1	11	0	0.333	0.9167	G _{q/11}
R. [RK]. {0,10}K[AGS][AGS]	1	11	0	0.333	0.9167	G _{q/11}
[ILV]A. {2,4}S. [ILV]	1	11	0	0.333	0.9167	G _{q/11}
[AGS]. [ILV]. {2,10}L. [FWY]	0	11	1	0.333	0.9167	G _{q/11}
[AGS][FWY]. . [FWY]	1	11	0	0.333	0.9167	G _{q/11}
S.S. {1,11}L.S	0	11	1	0.333	0.9167	G _{q/11}
[ILV]. L. {6,11}A. T	1	11	0	0.333	0.9167	G _{q/11}
K. {0,3}N.P	1	11	0	0.333	0.9167	G _{q/11}
[ILV]. L. {6,10}A. T	0	11	0	0.333	1	G _{q/11}
[RK][FWY]. . . .K	2	13	0	0.394	0.8667	G _{q/11}
[AGS]. S. [RK]. {2,10}F	1	13	0	0.394	0.9286	G _{q/11}
[ILV]. {3,6}S. Q	3	18	3	0.545	0.7500	G _{q/11}
C. [FWY]. {2,11}K	0	10	1	0.303	0.9091	G _{q/11}
C. [FWY]. {2,12}K	0	10	1	0.303	0.9091	G _{q/11}
S. . . [RK]A. {3,10}S	1	10	0	0.303	0.9091	G _{q/11}

loops, since from prior biochemical investigations it is unlikely this would be sufficient to provide an effective and selective G protein interaction (Wess 1998). Similarly whether additional predictive power can be gained from analysis of where the patterns match in the context of a particular intracellular domain, and their distance from the membrane remains to be investigated. Receptor-G protein recognition is known to be regulated by both post-transcriptional and post-translational modifications, likely of both the GPCR and the G-protein heterotrimer (Wess 1998). Analysing just the translated receptor coding sequence does not allow us to model such events. In spite

of these issues, the discovered patterns are sensitive and selective, enabling our construction of a useful predictor, allowing us to address a problem that has repeatedly been stated to be rather a difficult one (Wess 1998; Sautel and Milligan 2000; Horn, van der Wenden et al. 2000).

In conclusion this work combines protein sequence data with information not currently found in annotated sequence databases, or specialist companion databases such as GPCRDB (Horn, Vriend et al. 2001) to find patterns that can be used for making functional inferences concerning GPCR signalling. The patterns have a wide range of application, from predicting the G protein cou-

Table 3. List of generated patterns found to best represent a specific coupling mechanism. Column 1 shows the pattern as a regular expression, columns 2-4 show the number of matches to the different receptor-G protein coupling groups, columns 5 and 6 show sensitivity and specificity, respectively.

Pattern	G _{io}	G _{q/11}	G _s	Sensitivity	Specificity	Best class
Total	55	33	25			
A[ILV].{1,5}Y..[ILV].T	0	0	10	0.400	1	G _s
A.{1,5}RY...T	0	0	10	0.400	1	G _s
I...RY.{1,10}R	0	0	9	0.360	1	G _s
I...RY.{4,6}T	0	0	9	0.360	1	G _s
LR.{1,9}T...[ILV]	0	0	9	0.360	1	G _s
RS.{3,13}C[AGS]	0	0	9	0.360	1	G _s
[ILV].[FWY]H.{1,3}I	0	0	9	0.360	1	G _s
F.{1,4}Y...T	0	0	9	0.360	1	G _s
I...RY.{4,4}T	0	0	9	0.360	1	G _s
I...R[FWY]	0	0	9	0.360	1	G _s
I...RY...T	0	0	9	0.360	1	G _s
I...RY	0	0	9	0.360	1	G _s
[FWY].A.{2,6}Y..[ILV]	0	0	9	0.360	1	G _s
I.[AGS].{1,10}S...R	0	0	8	0.320	1	G _s
[ILV].[FWY]H.{3,12}T	0	0	8	0.320	1	G _s
L..H.[ILV]	0	0	8	0.320	1	G _s
[ILV].[FWY]H.[ILV]	0	0	8	0.320	1	G _s
[ILV].[FWY]H.I	0	0	8	0.320	1	G _s
A...[RK][RK]I	0	0	8	0.320	1	G _s
[AGS].{0,10}L..H.[ILV]	0	0	8	0.320	1	G _s
[ILV].[FWY]H.{3,10}T	0	0	8	0.320	1	G _s
[FWY]H.I.{0,3}T	0	0	8	0.320	1	G _s
S.{5,12}S.L.[RK]	0	0	8	0.320	1	G _s
S.{5,9}S.L.[RK]	0	0	8	0.320	1	G _s
Q.{0,9}S.L.[RK]	0	0	8	0.320	1	G _s
A.{1,5}RY..[ILV].T	0	0	8	0.320	1	G _s
F.{1,10}A...H	0	0	8	0.320	1	G _s
[ILV]..H.[ILV].{1,3}T	0	0	8	0.320	1	G _s
[FWY]H.I.{0,10}V	0	0	8	0.320	1	G _s
A..[FWY].{0,3}H	0	1	10	0.400	0.9091	G _s
I...[RK]Y.{4,6}T	0	0	10	0.400	1	G _s
A.{1,5}R[FWY]...T	1	0	10	0.400	0.9091	G _s
A.{2,6}Y..[ILV].T	0	1	10	0.400	0.9091	G _s
A..[FWY].{0,8}H	1	1	11	0.440	0.8462	G _s
[AGS].{1,5}RY...T	0	2	11	0.440	0.8462	G _s
I...[RK]Y.{1,10}R	0	1	9	0.360	0.9000	G _s
R[FWY]H.{5,14}R	0	1	9	0.360	0.9000	G _s
[RK]S.{3,13}C[AGS]	1	0	9	0.360	0.9000	G _s
[RK].[ILV].C.R	1	0	9	0.360	0.9000	G _s
[RK].[ILV].C.[RK]	1	0	9	0.360	0.9000	G _s

pling preferences of newly cloned receptors, to designing biochemical experiments to increase our understanding of the molecular basis of receptor-G protein coupling. Developing and improving ways to make such functional predictions between interacting proteins, and the subsequent reconstruction of cellular pathways, constitutes one of the key challenges to bioinformatics as we enter the post-genomic era.

ACKNOWLEDGEMENTS

The authors wish to thank Anders Krogh and Henrik Nielsen for the collaborative development of 7TMHMM.

MDRC would like to thank Rolf Apweiler for the generous provision of resources at the European Bioinformatics Institute.

REFERENCES

- Attwood, T. K., M. D. R. Croning, et al. (2000). PRINTS-S: the database formerly known as PRINTS. *Nucl. Acid Res.*, **28**(1), 225-227.
- Bairoch, A. and R. Apweiler, (2000). The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucl. Acid Res.*, **28**(1), 45-48.
- Birney, E. (2001). www.ensembl.org.

Table 4. This table presents the predictions for 10 sequences that are unrelated to the training set. Column 1 lists the SWISS-PROT or TrEMBL accession numbers, column 2 the predicted receptor-G protein coupling. Column 3 shows two numbers, the total number of matches and the number of matches contributed by the pairs and triplets to the predicted class. Column 4 shows the protein's description.

Accession	Class	Hits (class/total)	Protein description
P49190	G _s	123 / 124	PARATHYROID HORMONE RECEPTOR
Q03431	G _s	123 / 132	PARATHYROID HORMONE/PARATHYROID HORMONE-RELATED PEPTIDE RECEPTOR
Q02643	G _s	123 / 124	GROWTH HORMONE-RELEASING HORMONE RECEPTOR
O95838	G _s	181 / 192	GLUCAGON-LIKE PEPTIDE 2 RECEPTOR
P41180	G _{q/11}	11 / 14	EXTRACELLULAR CALCIUM-SENSING RECEPTOR
P47872	G _s	123 / 124	SECRETIN RECEPTOR
P43220	G _s	125 / 142	GLUCAGON-LIKE PEPTIDE 1 RECEPTOR
P48546	G _s	124 / 140	GASTRIC INHIBITORY POLYPEPTIDE RECEPTOR
P25105	G _{q/11}	4 / 7	PLATELET ACTIVATING FACTOR RECEPTOR
O43613	G _{q/11}	52 / 56	OREXIN RECEPTOR TYPE 1

- Bockaert, J. and J. P. Pin (1999). Molecular tinkering of G protein-coupled receptors: an evolutionary success. *EMBO J.*, **18(7)**, 1723-1729.
- Bourne, H. R. (1997). How receptors talk to trimeric G proteins. *Curr. Opin. Cell Biol.*, **9(2)**, 134-142.
- Brazma, A., I. Jonassen, et al. (1998). Approaches to Automatic Discovery of Patterns in Biosequences. *Journal of Computational Biology*, **5(2)**, 277-304.
- Fleischmann, W., S. Möller, et al. (1999). A novel method for automatic functional annotation of proteins. *Bioinformatics*, **15(3)**, 228-233.
- Gether, U. (2000). Uncovering molecular mechanisms involved in activation of G protein-coupled receptors. *Endocrine Reviews*, **21(1)**, 90-113.
- Horn, F., G. Vriend, et al. (2001). Collecting and harvesting biological data: The GPCRDB and NuclearRDB information systems. *Nucl. Acid Res.*, **29(1)**, 346-349.
- Horn, F., E.M. van der Wenden, et al. (2000). Receptors coupling to G proteins: is there a signal behind the sequence? *Proteins.*, **41(4)**, 448-459.
- LeVine III, H. (1999). Structural features of heterotrimeric G-protein-coupled receptors and their modulatory proteins. *Mol. Neurobiol.*, **19(2)**, 111-149.
- Liu, J. and J. Wess (1996). Different single receptor domains determine the distinct G protein coupling profiles of members of the vasopressin receptor family. *J. Biol. Chem.*, **271**, 8772-8778.
- Livingstone, C. D. and G. J. Barton (1993). Protein sequence alignments: a strategy for the hierarchical analysis of residue conservation. *Comp. Appl. Biosci.*, **9(6)**, 745-756.
- Möller, S., M. D. R. Croning, et al. (2001). Evaluation of Methods for the prediction of membrane spanning regions in transmembrane proteins. *Bioinformatics*, **in press**.
- Möller, S., M. D. R. Croning, et al. (2001). 7TMHMM. *to be submitted*
- Möller, S., U. Leser, et al. (1999). EDITtoTrEMBL: a distributed approach to high-quality automated protein sequence annotation. *Bioinformatics*, **15(3)**, 219-227.
- Sautel, M. and G. Milligan (2000). Molecular manipulation of G-protein-coupled receptors: a new avenue into drug discovery. *Current Medicinal Chemistry*, **7(9)**, 889-896.
- Scherf, M., A. Klingenhoff, et al. (2000). Highly specific localization of promoter regions in large genomic sequences by PromoterInspector: a novel context analysis approach. *J Mol. Biol.*, **297(3)**, 599-606.
- Schoneberg, T., G. Schultz, et al. (1999). Structural basis of G protein-coupled receptor function. *Mol. Cell. Endocrinol.*, **151(1-2)**, 181-193.
- Simon, M. I., M. P. Strathmann, et al. (1991). Diversity of G proteins in signal transduction. *Science*, **252(5007)**, 802-808.
- TiPS (2000). *Receptor & ion channel nomenclature supplement*.
- Vilo, J. (1998). *Discovering Frequent Patterns from Strings*. Department of Computer Science, University of Helsinki.
- Vilo, J., A. Brazma, et al. (2000). Mining for putative regulatory elements in the yeast genome using gene expression data. *ISMB*, **8**, 384-394.
- Wess, J. (1998). Molecular basis of receptor/G-protein-coupling selectivity. *Pharmacol. Ther.*, **80(3)**, 231-264.
- Wilson, S., D. J. Bergsma, et al. (1998). Orphan G-protein-coupled receptors: the next generation of drug targets? *Br. J. Pharmacol.*, **125(7)**, 1387-1393.