Mining of assembled expressed sequence tag (EST) data for protein families: Application to the G protein-coupled receptor superfamily

Darrell Conklin, David P. Yee, Robert Millar, Jacob Engelbrecht and Henrik Vissing

Date received (in revised form): 11th November 1999

Darrell Conklin
is a computational biologist at ZymoGenetics. His research is focused on techniques for protein fold recognition and DNA sequence assembly with the goal of discovery of potential therapeutic proteins.

David Yee
is a senior scientist at Millennium Pharmaceuticals, Inc. He is interested in the application of computational techniques to address biological problems in the areas of gene discovery and functional prediction.

Robert Millar
is Director of the MRC Reproductive Biology Unit in Edinburgh. His research interests are on g protein-coupled receptors.

Jacob Engelbrecht
has researched in the field of gene prediction and sequence structure. As a research scientist in the Scientific Computing Department of Novo Nordisk, he is responsible for implementing bioinformatics tools.

Henrik Vissing
is Section Head of Bioinformatics and Genomics in Novo Nordisk and Project Manager of the Health Care Discovery Bioinformatics project. His area of research is in the field of drug target identification and validation.

Keywords: *bioinformatics, database mining, G protein-coupled receptor, expressed sequence tag, sequence assembly, rhodopsin family*

Henrik Vissing,
Novo Nordisk A/S,
Novo Alle,
Bagsvaerd,
DK-2880,
Denmark

E-mail: hv@novo.dk

Abstract

The availability of large expressed sequence tag (EST) databases has led to a revolution in the way new genes are identified. Mining of these databases using known protein sequences as queries is a powerful technique for discovering orthologous and paralogous genes. The scientist is often confronted, however, by an enormous amount of search output owing to the inherent redundancy of EST data. In addition, high search sensitivity often cannot be achieved using only a single member of a protein superfamily as a query. In this paper a technique for addressing both of these issues is described. Assembled EST databases are queried with every member of a protein superfamily, the results are integrated and false positives are pruned from the set. The result is a set of assemblies enriched in members of the protein superfamily under consideration. The technique is applied to the G protein-coupled receptor (GPCR) superfamily in the construction of a GPCR Resource. A novel full-length human GPCR identified from the GPCR Resource is presented, illustrating the utility of the method.

Introduction

The large public and corporate databases of ESTs (expressed sequence tags)[1] contain a wealth of information on both known and unknown expressed genes. EST sequencing is an efficient way to 'tag' a large number of expressed genes. ESTs are obtained by short (roughly 300 nucleotides) reads of clone ends from selected cDNA libraries. EST databases continue to be a rich source for the cloning of novel members of gene families.

The problem of high redundancy in EST databases is now well understood.[2–5] A powerful way to manage this redundancy is to assemble clusters of ESTs representing the same message into longer virtual cDNA sequences.[6–8] There are several advantages to working with assemblies rather than individual ESTs: first, there are fewer sequences to analyse; second, the assembled sequences are longer and potentially contain more interpretable coding sequence than their individual component ESTs; third, sequencing errors present in individual ESTs may be corrected during the assembly process; fourth, the virtual cDNA sequences may extend to the 5' end of the mRNA, greatly facilitating cloning of the gene in the laboratory.

A protein superfamily is a set of sequences having a similar structure but not necessarily exhibiting significant sequence similarity. Given a protein superfamily, rigorously specified search techniques are necessary to 'mine' assembled EST databases for all superfamily members. The computational biologist is confronted by the often conflicting goals of high search sensitivity (identifying all true positives) versus high search specificity (permitting few false positives). In this paper, we use a simple yet effective technique for database mining of large superfamilies, where it is assumed that the known superfamily members cover a wide area of the actual sequence space. In operational terms, it is assumed that any new superfamily member will have a significant pairwise alignment score with at least one known superfamily member. Results from individual searches of all superfamily members are integrated into one list. Search sensitivity arises by choosing a suitably low threshold score for each search. Search selectivity arises from the use of a pruning procedure that aims to remove false positives from the integrated search results.

G protein-coupled receptors

The G protein-coupled receptors (GPCRs) form a large superfamily of proteins that transduce signals across the cell membrane. This signalling is performed through the activation of cytosolic messengers called G proteins. Recently GPCRs have been shown to couple also to an Na^+/H^+ exchanger regulatory factor, and to c-Src (via β-arrestin) without G proteins as mediators.[9] The structural hallmark of the GPCR superfamily is seven hydrophobic transmembrane domains connected by hydrophilic loops. GPCRs are a rich source of targets for drug discovery.[10]

Web resources for GPCRs

Several web-based resources for the inspection and analysis of GPCR sequences exist; only a few most relevant to this study are listed. The Swiss-Prot 7-transmembrane receptor list (http://www.expasy.ch/cgi-bin/lists?7tmrlist.txt) annotates and classifies GPCRs into functional categories. The GCRDb database (http://www.gcrdb.uthscsa.edu/) classifies GPCR sequences into one of six broad families, and also provides for retrieval by organism and ligand type. The GPCR database (GPCRDb) (http://swift.embl-heidelberg.de/7tm/) contains useful derived data such as phylogenetic trees and snake-like plots illustrating individual receptor topologies.

In this paper a technique is described for handling the large amount of available EST data representing known and potentially novel receptors. The goal of the system is not to attain maximum selectivity but rather to produce a set of assembled EST data enriched in superfamily members. To illustrate the utility of the technique, a novel GPCR, with only weak homology to other members of the GPCR superfamily, is presented.

METHODOLOGY

REX/ACE

A collection of approximately 7.5 million ESTs from public and proprietary databases representing various tissue sources and organisms was assembled using the REX (Recursive EST eXtender) algorithm. The result is called the ACE (Assembled Consensus EST) database.[7] ACE is maintained using a relational database (Sybase Inc., Emeryville, CA). Information in ACE includes the consensus sequence, the anchor sequence, the organism, the date of assembly and library information for every component EST sequence.

The GPCR query set

From the Swiss-Prot database, release 37.0, the 1,073 entries in the GPCR section were extracted. From this set all sequences that were of human origin, or did not have a human orthologue, were extracted into the query set. In this manner, for example, the human adenosine A1 receptor AA1R_HUMAN was extracted but not its mouse orthologue AA1R_MOUSE. This produced a query set of 423 protein sequences, of which 208 were of human origin.

Searching for homologues and integrating results

The tblastn program,[11] the gapped alignment implementation WU-BLAST

Integrating multiple query results

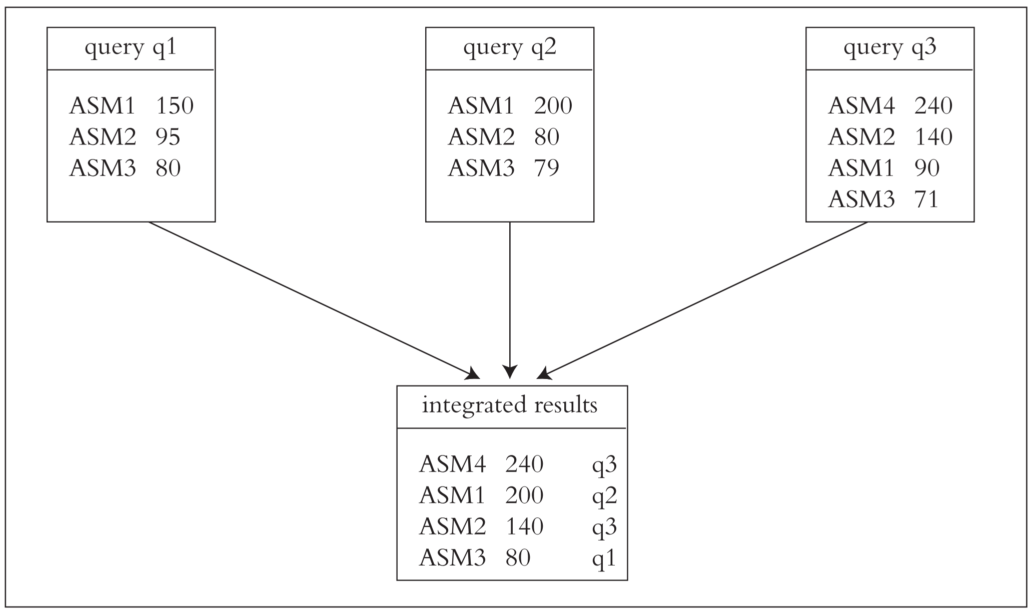


Figure 1: A simple artificial example of the multiple query integration method. Query sequences q1, q2 and q3 match the assemblies ASM1, ASM2, ASM3 and ASM4 above the score threshold of 70. The score of an assembly in the integrated results list is that of its highest-scoring query

***TBLASTN* parameters for queries**

2.0, with default parameters except with S = 70, V = 10,000, B = 10,000, filter = seg + xnu, was run with each query in the query set against the ACE database. Query sequences were masked in areas of translated ALU repeats and other translated repetitive DNA sequences (eg the thromboxane A2 receptor TA2R_HUMAN has translated ALU sequence at its C-terminus). Computing hardware used was an SGI Origin 2000 with 32 processors and 4GB of shared memory. Results for every query were integrated together into a single non-redundant list. This is done by a simple algorithm which produces a list of query/assembly pairs with every assembly associated with its highest scoring query (see Figure 1).

Pruning false positives

Candidate assemblies are back-searched against full Swiss-Prot database

The individual tblastn searches may be overly sensitive and some assemblies may be readily identifiable as members of other protein families. In addition, some GPCRs contain common protein domains in their extracellular portion (eg epidermal growth factor (EGF) domains, leucine-rich repeats). A simple measure to remove potential false positive assemblies was undertaken. For each candidate assembly, the blastx program[11] was used to 'back search' the assembly against the full Swiss-Prot database. If the highest scoring Swiss-Prot sequence was a GPCR (as determined by its membership in the Swiss-Prot GPCR section), the assembly was retained in a list of putative GPCR superfamily members. Note that the highest-scoring blastx hit need not be the same as the highest-scoring query for the assembly: it is only required that they are both GPCRs.

Building the GPCR Resource

Each candidate assembly passing the false positive test forms an entry in the final GPCR Resource, which is coded in HTML to be used by a web browser. Assemblies are classified into one of the 36 Swiss-Prot GPCR subfamilies (eg adrenergic receptors, releasing hormone receptors). Each entry contains the assembly name, the highest-scoring Swiss-Prot name, the blastx score and the percentage identity. Various fields are linked to further web pages. The score is a link which leads to a page containing a tfastx3[12] alignment between the assembly and the Swiss-Prot query. In addition to the tfastx3 alignment, this page contains buttons that allow a database miner to annotate the assembly as identical to a known sequence or potentially interesting as a novel gene. Clicking on an assembly name gives full information on all component ESTs in the assembly, their tissue source and the nucleotide sequence of the assembly.

Assemblies span a wide range of GPCR categories

The GPCR Resource containing 855 human assemblies

145	Orphan receptors
112	Odorant/olfactory and gustatory receptors
110	Family 2 (B) receptors
47	Adenosine and adenine nucleotide receptors
38	Family 3 (C) receptors (metabotropic glutamate and calcium receptors)
37	Opsins
33	Adrenergic receptors
30	Neuropeptide Y receptors
28	Serotonin receptors
23	Chemokines and chemotactic factors receptors
20	Prostanoids receptors
18	Glycoprotein hormones receptors
16	Dopamine receptors
14	Melanocortins receptors
14	Endothelin receptors
13	Viral receptors
13	Other receptors
12	Somatostatin receptors
11	Tachykinin receptors
10	Vasopressin / oxytocin receptors
10	Releasing hormones receptors
10	Melanotonin receptors
10	Bombesin receptors
10	Acetylcholine (muscarinic) receptors
9	Opioid peptides receptors
9	Family 4 receptors (fungal receptors)
9	Cannabinoids receptors
8	Proteinase-activated receptors
7	Neurotensin receptors
6	Family 5 receptors (slime mould receptors)
6	Angiotensin receptors
5	Archebacterial receptors
4	Histamine receptors
4	Bradykinin receptors
2	Platelet activating factor receptors
2	Cholecystokinin / gastrin receptors

Figure 2: Numbers of entries in each Swiss-Prot category for the human assemblies in the GPCR Resource

RESULTS

The GPCR Resource

The ACE database of assembled ESTs comprises approximately 650,000 assemblies, of which 270,000 are non-singleton (ie containing more than one EST) assemblies. All assemblies are organism-specific: constructed only from component ESTs from the same organism. Within the set of non-singleton assemblies, 213,000 are human, 45,000 are mouse and the remainder are from other organisms. As outlined in Methodology, the ACE database was queried with non-orthologous GPCR sequences and the results were integrated. This produced a list of 7,843 distinct query/assembly pairs representing candidate GPCR superfamily members. After pruning this list to remove potential false positives, 1,153 unique assemblies remained. Some 855 of the assemblies were human; the remainder were assemblies of mouse and rat ESTs. Of the human assemblies, 312 were singletons. The 855 human assemblies had a mean length of 970 nucleotides, and together they covered a total of 23,742 component ESTs. Each assembly was classified into a Swiss-Prot GPCR category: Figure 2 shows the number in each category for the human assemblies. Figure 3 shows a histogram of the blastx percentage identity for the human assemblies in the GPCR Resource. Two areas of the histogram are apparent: one at about 70 per cent and higher representing putative identities or orthologues to receptors in Swiss-Prot, and another with a peak around 35 per cent representing putative novel receptors.

A novel GPCR

By scanning the GPCR Resource, a human assembly indexed to the Swiss-Prot category of adenosine and adenine nucleotide receptors was identified. The assembly was associated with the chicken purinergic receptor 5

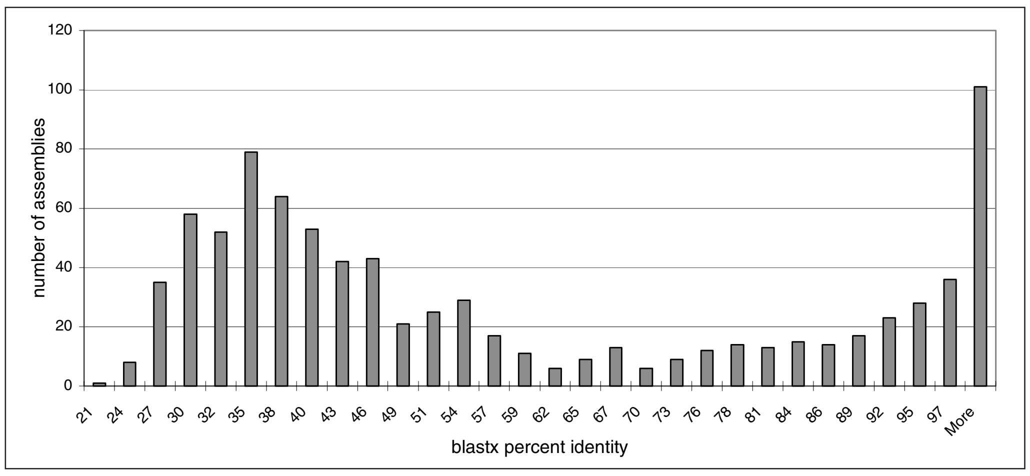
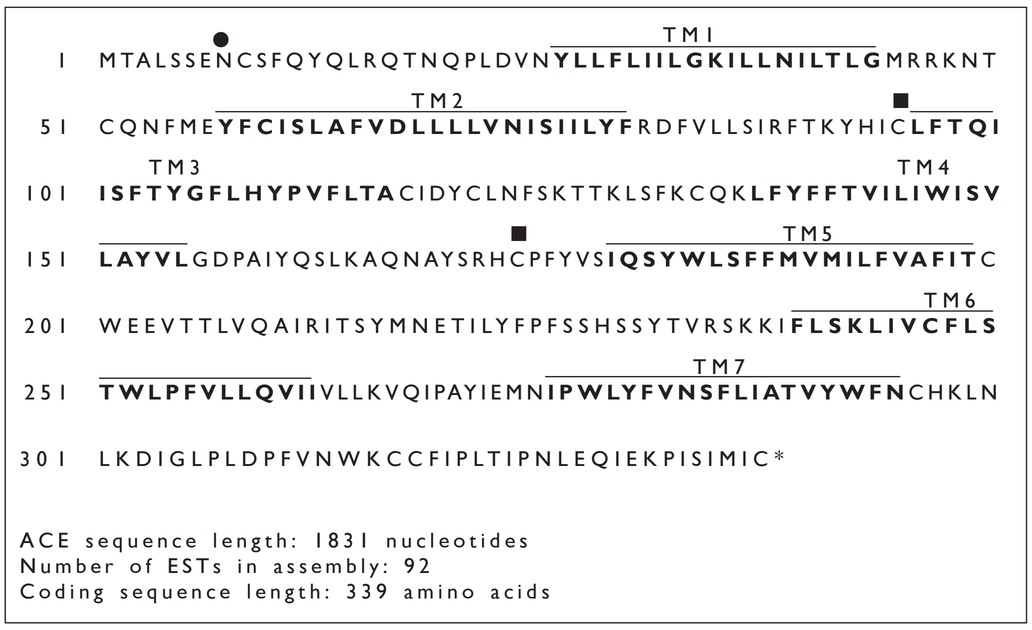


Figure 3: Histogram representing the blastx percentage identity of an assembly to its closest GPCR, for all human assemblies in the GPCR Resource


●
TM1
■
TM2
TM3
TM4
■
TM5
TM6
TM7

1 MTALSSENCSFQYQLRQTNQPLDVN**YLLFLIILGKILLNILTLG**MRRKNT

51 CQNFME**YFCISLAFVDLLLLVNISIILYF**RDFVLLSIRFTKYHIC**LFTQI**

101 **ISFTYGFLHYPVFLTA**CIDYCLNFSKTTKLSFKCQK**LFYFFTVILIWISV**

151 **LAYVL**GDPAIYQSLKAQNAYSRHCPFYVS**IQSYWLSFFMVMILFVAFIT**C

201 WEEVTTLVQAIRITSYMNETILYFPFSSHSSYTVRSKKI**FLSKLIVCFLS**

251 **TWLPFVLLQVII**VLLKVQIPAYIEMN**IPWLYFVNSFLIATVYWFN**CHKLN

301 LKDIGLPLDPFVNWKCCFIPLTIPNLEQIEKPISIMIC*

ACE sequence length: 1831 nucleotides
Number of ESTs in assembly: 92
Coding sequence length: 339 amino acids

Figure 4: A novel human receptor identified in the GPCR Resource, called GPCR150, showing the sequence with the predicted transmembrane domains underlined, and the conserved cysteine pair indicated (■). Also indicated is a putative N-linked glycosylation site at the N-terminus (●)

GPCR: a novel GPCR identified in the GPCR Resource

(P2Y5_CHICK) at a blastx score of 112 and only 22 per cent identity. Subsequent analysis of an open reading frame in the assembly indicated a putative full-length gene without any apparent frameshift errors. The assembly was 1,831 nucleotides in length, and comprised 92 ESTs, derived from a wide variety of tissue sources. To verify the sequence, the cDNA was isolated by RT-PCR (reverse transcription polymerase chain reaction) using a gene-specific oligo designed to the IMAGE consortium[13] clone #1420265; the isolated PCR product was named *GPCR150* (Genbank accession number AJ249248). Figure 4 shows the polypeptide sequence of GPCR150 with the TMHMM[14] (transmembrane hidden Markov model) transmembrane helix predictions underlined. Although the novel receptor has homology with the purinergic receptors and shares a number of conserved amino acids and motifs,[15,16] it also differs from the rhodopsin family in a number of unusual ways.

GPCR150 has the conserved N in TM1 (nomenclature: EC1 connects TM2 and TM3; IC1 connects TM1 and TM2) and the LXXXD motif in TM2. It also has a C at the C-terminal end of EC1 and its potential partner for disulphide bridge formation in EC2 as in all GPCRs. The conserved W in TM4 is also present as is the characteristic long IC3 with putative phosphorylation sites which are important in receptor uncoupling, specificity of G protein coupling and receptor internalisation.[9] The motif FXXXWXP, characteristic of rhodopsin GPCRs, occurs in TM6. The NPXXY of TM7 lacks the N and is positioned at the beginning of TM7 instead of the end as in the rhodopsin family of receptors. The C-terminal tail has characteristic putative phosphorylation and palmitoylation sites. Surprisingly, the crucial [DE]RXXX[IV]XXP motif following TM3 is totally absent from GPCR150. The R residue in this motif forms a salt bridge with the acidic residue, with this salt bridge disrupted in the activated receptor.[17] It is possible that GPCR150 is activated in another manner.

Motifs in the GPCR 150 sequence

DISCUSSION

This paper has outlined a general technique for mining of assembled EST data in protein superfamilies that have a natural functional or evolutionary subclassification. Searches are performed for all queries in the superfamily and the results integrated into one list and sorted by decreasing score. The score of a particular assembly in this list is that of its highest-scoring query sequence. A similar scheme for protein family searching has been reported,[18] where the results are sorted by the average, rather than maximum, score over all queries. It was demonstrated that this simple technique performed equally as well at separating true from false positives as more 'advanced' techniques such as profile and hidden Markov models. For the GPCRs, the average score method is not likely to be more effective than the maximum score method, given the large sequence space covered by the GPCR superfamily. Another idea for integrating multiple query search results is to sort the integrated results by number of queries matching above a certain threshold score, choosing some cutoff (eg 5) for the number of matching queries.[19] This technique is not likely to be effective for the GPCRs given that putative novel GPCRs that have only one matching query have been identified.

Other techniques for multiple integration

Our GPCR Resource has identified a novel full-length human GPCR with low homology to purinergic receptors of the rhodopsin superfamily. However, GPCR150 is too distant to be placed in any specific family, since it has similar weak homology to the chemokine, serotonin, melatonin and purigenic receptor all in the rhodopsin family. Although this receptor has several of the hallmarks of the family, it lacks the crucial [DE]R motif at the cytosolic end of TM3. Since the overall homology is low, this receptor may represent a novel family of GPCRs or an isolated member of the rhodopsin family which is activated in another way.

The false positive pruning procedure, which relies on a 'back search' with an assembly against the full Swiss-Prot database, may be subject to occasional errors. The requirement that an occurrence of a common protein domain in an assembly is most similar to an occurrence in a GPCR may not always be valid, causing true positives to be discarded. Also, for assemblies having insignificant scores to any known protein, there is a small chance that the back search can indicate a match with a GPCR at random. This will cause the erroneous retention of a false positive in the database.

In the superfamily searching method reported here, the individual searches are performed using tblastn, and the false positive checks using blastx. Though assemblies of ESTs tend to have much higher quality than their component ESTs, some frameshifts cannot be corrected by assembly algorithms owing to poor sequence quality and lack of consensus. Since tblastn will not join high-scoring segment pairs across different reading frames, it is possible that a lack of sensitivity is introduced. At the expense of substantially longer computation times, it is possible to use the tfastx/fastx pair of searching tools,[12] which handle frameshifting, to construct the GPCR Resource.

The database mining technique can be applied to human genomic sequence data

A well-known limitation of EST databases is that non-subtracted cDNA libraries are a sample of messages reflecting their abundance in the tissue source, with the effect that rare messages will not have a high probability of being tagged until large numbers of cDNAs are sampled for sequencing.[2] In addition, many rare genes are expressed in very discrete cell types or developmental states, and these may not have formed a basis for cDNA library construction. Within a few years, all human genes will be present in raw human genomic sequence. Therefore we have initiated the development of a parallel GPCR Resource that applies the techniques described in this paper to a database of genes predicted from human genomic sequence. This technique will solve the expression level constraints of EST data.

Acknowledgements

We acknowledge the valuable assistance of Kate Hansen in the cloning of GPCR150, and Robert Bywater for discussions on G protein-coupled receptors.

References

1. Boguski, M. S., Lowe, T. M. and Tolstoshev, C. M. (1993), *Nature Genetics*, Vol. 4, pp. 332–333.
2. Wan, J. *et al.* (1996), *Nature Biotechnol.*, Vol. 14, pp. 1685–1691.
3. Hillier, L. *et al.* (1996), *Genome Res.*, Vol. 6, pp. 807–828.
4. Boguski, M. and Schuler, G. (1995), *Nature Genetics,* Vol. 10, pp. 369–371.
5. Wolfsberg, T. and Landsman, D. (1997), *Nucleic Acids Res.*, Vol. 25(8), pp. 1626–1632.
6. Burke, J., Wang, H., Hide, W. and Davison, D. (1998), *Genome Res.*, Vol. 8, pp. 276–290.
7. Yee, D. P. and Conklin, D. (1998), in 'Proceedings of the Sixth International Conference on Intelligent Systems for Molecular Biology', AAAI Press, Menlo Park, CA, pp. 203–211.
8. Sutton, G., White, O., Adams, M. and Kerlavage, A. (1995), *Genome Sci. Technol.*, Vol. 1(1), pp. 9–19.
9. Schwartz, T. and Ijzerman, A. (1998), *Trends Pharmacol. Sci.*, Vol. 19(11), pp. 433–436.
10. Stadel, J., Wilson, S. and Bergsma, D. (1997), *Trends Pharmacol. Sci.*, Vol. 19(11), pp. 430–437.
11. Altschul, S., Warren, G., Miller, W., Meyers, E. and Lipman, D. (1990), *J. Mol. Biol.*, Vol. 215, pp. 403–410.
12. Pearson, W., Wood, T., Zhang, Z. and Miller, W. (1997), *Genomics*, Vol. 46(1), pp. 24–36.
13. Lennon, G.G., Auffray, C., Polymeropolous, M. and Soares, M. B. (1996), *Genomics*, Vol. 33, pp. 151–152.
14. Sonnhammer, E., von Heijne, G. and Krogh, A. (1998), 'ISMB '98', AAAI Press, Menlo Park, CA, pp. 175–182.
15. van Rhee, A., Fischer, B. and Jacobson, K. (1995), *Drug Design Discovery*, Vol. 13, pp. 133–154.
16. Sawzdargo, M., George, S., Nguyen, T., Xu, S., Kolakowski, L. and O'Dowd, B. (1997), *Biochem. Biophys. Res. Commun.*, Vol. 239(2), pp. 543–547.
17. Ballesteros, J., Kitanovic, S.,Guarnieri, F., Davies, P., Fromme, B. J., Konvicka, K., Chi, L., Millar, R. P., Davidson, J., Weinstein, H. and Sealfon, S. C. (1998), *J. Biol. Chem.*, Vol. 273(17), pp. 10445–10453.
18. Grundy, W. (1998), 'RECOMB '98: Second Annual International Conference on Computational Molecular Biology', ACM Press, New York, pp. 94–100.
19. Pegg, S. and Babbitt, P. (1999), *Bioinformatics*, Vol. 15, pp. 729–740.