In silico identification of novel therapeutic targets

D. Malcolm Duckworth and Philippe Sanseau

The availability of the human genome sequence is a 'once in a lifetime' opportunity for scientists to uncover all possible human drug-targets. As the sequence is very large, the best way to identify new genes rapidly is by computational (*in silico*) methods. There are now many examples in which pharmaceutical companies have identified genes of interest initially by *in silico* analysis. High-throughput data-generation techniques, such as microarray analysis, are key to the generation of human genome data. Bioinformatics techniques are therefore certain to play an increasingly important role in drug discovery.

D. Malcolm Duckworth
and *Philippe Sanseau
GlaxoSmithKline
Discovery Bioinformatics
Europe
Gunnels Wood Road
Stevenage
UK SG1 2NY
tel: +44 1438 768 119
fax: +44 1438 764 231
*e-mail: ps14446@gsk.com

▼ The completion in 2001 of the draft sequence of the human genome is clearly one of the major achievements of modern biology [1,2]. This very large data source contains, in theory, all possible human targets for therapeutic intervention and must be exploited. Active computational analysis to identify novel therapeutic opportunities is now in progress.

In silico mining of human sequence data

As of March 2002, 53 microbial genome sequencing projects have been completed (see http://www.tigr.org/tdb/mdb), and this sequence information is being actively mined to identify antibacterial targets for therapeutic intervention. However, this review focuses mainly on *in silico* analysis of human data.

Identifying the exact number of human genes is still a significant endeavour for genomics researchers, and estimates range from ~28,000 to >100,000. A recent publication detailing the finished sequence of chromosome 20 extrapolates the data to give a total of ~31,500 genes for the whole genome [3], which is in agreement with initial publications [1,2]. Another recent article estimates the gene number to be between 41,000 and 45,000 [4]. However, is has been suggested that current predicted numbers be treated with caution [5].

Not all human genes have the potential to become successful drug targets, and although ~500 form the basis of current drug therapies [6,7], the potential number of targets could, in fact, be 10-fold greater [8]. Most of the 500 or so targets can be classified into several key protein-families. Approximately 45% are G-protein-coupled receptors (GPCRs), and some well-known drugs, such as beta blockers, anti-psychotics and histamine antagonists, are directed at receptors in a subclass of this family [6,7]. Enzymes (including proteases and kinases) represent a further 28% of the total [6,7]. The remainder includes other important families such as nuclear receptors and ion channels. As these classes of genes have been successfully exploited in the past – mostly by small-molecule intervention (so-called drugable targets) – the identification and characterization of new family members is of great interest to the pharmaceutical industry.

Data sources

Two principal sequence data sources – expressed sequence tags (ESTs) and genomic sequence – have been used to mine for new human genes. Expressed sequence tags were widely used in the 1990s [9], and proved successful in the identification of targets such as cathepsin K [10] and the orexin receptors [11]. Some companies have based their drug discovery efforts on identifying potential new targets from EST collections; for example, Human Genome Sciences (Rockville, MD, USA) has been very active in this field [12]. Expressed sequence tags are now used in combination with the draft human genome sequence to uncover transcripts. Compared with EST, genomic sequence offers (in theory) higher sequence quality and access to all transcripts, gene structures and regulatory regions. The disadvantages of the genomic draft are that the data are still

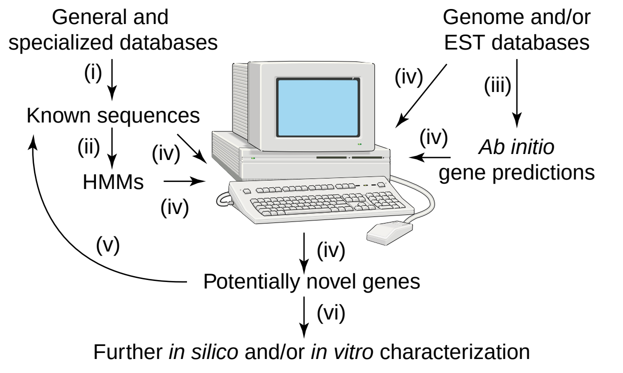
1359-6446/02/$ – see front matter ©2002 Elsevier Science Ltd. All rights reserved. PII: S1359-6446(02)02282-1

dynamic and fragmented. Unlike ESTs, genomic sequence offers no information on tissue expression. More details of both sequence data sources are discussed elsewhere [13].

Methods

Several different *in silico* approaches can be used to identify new family members of known target classes. Homology searching remains the method of choice for computational identification of transcripts. In this instance, known sequences are used as seeds, as described in Fig. 1. The first step is to identify relevant sequences from the wide variety of data sources; sequence databases such as GenBank are a good general source of data [14]. More-specialized databases can be used for GPCRs [15,16], ion channels [17], proteases [18], nuclear receptors [19] or kinases [20]. Although human gene sequences are the preferred input for searching, sequences derived from other species, such as mouse, *Drosophila* or *Caenorhabditis elegans*, can also be useful probes. However, one should not expect comparable numbers of gene family members across species, and thus identifying orthologues (the same gene in different species assumed to have the same function) is not always easy. In the extreme, orthologues might not exist at all. For example, although the number of GPCRs in the human genome is estimated to be ~700 [21], the number of receptors in *C. elegans* is closer to 1100 [22]. Similar differences are also found for other target classes, such as nuclear receptors, in which *C. elegans* has >250 compared with ~50 in humans [23].

In most cases, known amino acid sequences are chosen over nucleotide sequences because of their conservation. If common domains are present in the gene, specific conserved motifs can be used for searching. Use of the full gene sequence is more likely to identify unrelated members because of the possible presence of non-conserved regions. In addition, some genes can have dual functions indicated by different domains. For example, the recently cloned ion channel LTRPC7 also has a kinase domain [24]. Use of conserved domains is therefore preferred on most occasions. It is recommended that known sequences of different target classes are stored electronically for ease of practice. Two principal approaches can be used once the known sequences have been selected. The simplest method is a direct, amino acid homology-based search against the genome and ESTs sequence databases, using programs such as BLAST [25], Procrustes [26] or Genewise [27]. The second approach uses gene prediction programs, as a first step, to find *ab initio* novel genes from the human genome sequence. Many of these algorithms use statistical approaches, such as codon usage, amino acid usage or periodicities in coding regions, and profiles of splicing signal predictions, to predict genes. The most popular of these programs are Genscan [28] and Grail [29]. In a second step, the amino acid sequences for the *ab initio* predicted genes are compared by homology searches with known genes, using BLAST, as in the first approach. More-sensitive BLAST programs, such as PSI-BLAST, have been developed to identify more distant members of a gene family [30]. In addition, Hidden Markov Models (HMMs) from known genes can be compared with *ab initio* predicted transcripts [31]. Hidden Markov Models are based on probabilities and enable searches that use a multiple alignment. A library of HMMs representing all proteins of known structure has been made available and can be used for homology searches [32]. All novel genes identified *in silico* should be included in a new cycle of searches to look for related family members.


General and specialized databases
(i)
Known sequences
(ii)
HMMs
(iv)
(iv)
(v)
Genome and/or EST databases
(iv)
(iii)
Ab initio gene predictions
(iv)
(iv)
Potentially novel genes
(vi)
Further *in silico* and/or *in vitro* characterization

DDT: InfoBiotech supplement

Figure 1. *In silico* identification of novel genes. Known amino acid sequences from the target families of interest are selected from generalized or specialized databases (i) (see main text for examples). These sequences can be human or derived from other species to identify orthologues. In a second step, Hiddden Markov Models (HMMs) can be built (ii). Genes can be predicted by *ab initio* methods (e.g. Genscan) from genome databases (iii). (iv) Potential novel genes are identified by comparing sequences from the outputs of steps (i) or (ii) from databases of genomic sequence or expressed sequence tags, or with the output of step (iii). For homology searching, programs such as BLAST or FASTA can be used. Novel genes can feed another cycle of the same process to extend the search to more distant family members (v). Potential novel transcripts are further characterized, for example, by using ESTs to help verify the gene structure and provide confirmatory evidence that the sequence can be expressed (vi). Functional domains can also be identified by sequence analysis. *In vitro* work then is then rapidly initiated, and starts by cloning the corresponding cDNA to confirm the sequence established *in silico*.

Methods that directly identify novel multi-transmembrane (TM) proteins from genomic databases are less commonly used. However, one example is an algorithm developed to identify novel GPCRs using structural features by statistically characterizing the different physicochemical properties of known transmembrane proteins [33]. Its utility was then demonstrated by identifying novel GPCRs from the *Drosophila* genome.

In many cases, the sequence accuracy of novel potential transcripts is limited. It is therefore necessary to refine the predictions by further *in silico* analysis. For example, combinations of gene prediction programs can be applied, and gene structures can be compared with known transcripts. If ESTs are available, they can be clustered to delineate the transcript, especially at the 5′ and 3′ ends of the gene, or analysed to give information on splice variants [34]. Mouse and human cross-species genomic sequence analysis can be used to identify conserved regions; this is particularly useful in the identification of both 5′ untranslated regions of genes and starting methionines of the coding transcript.

The Ensembl genome database project (http://www.ensembl.org) is a public initiative providing stable, automatic annotation of the human genome sequence, and pre-computes some of the steps described previously [35]. For example, it includes in the gene building system, various *ab initio* gene prediction methods, homology searches and HMMs. Furthermore, EST data are also being integrated into the gene annotation system. As these annotations are available as an interactive web service, it is convenient and straightforward for the end-user to query the database in different ways. The website provides alternative views of the data that can be selected by the user. Interestingly, it is also possible to integrate a user's own annotations or data sources into Ensembl.

It should be noted, however, that no single method can be chosen for *in silico* identification of novel genes. In most cases, a combination of approaches will be the best way to identify all possible members of a particular gene family.

As GPCRs represent the largest class of known drug targets, with applications to a range of therapeutic conditions (e.g. inflammation, obesity and cardiovascular diseases), we will use this target class to illustrate the importance of bioinformatics in the process of identifying *in silico* novel targets, focussing particularly on issues surrounding the validation of these targets in disease intervention. Similar methods can be applied to other families of drug targets.

Predicting novel GPCRs *in silico*

What are GPCRs?

GPCRs are cell-surface molecules involved in the regulation of different physiological processes, such as synaptic transmission in the central nervous system, regulation of the immune system, smell, taste, and so on. Ligand binding from the extracellular surface to the GPCR promotes a conformational change and stimulates intracellular nucleotide-binding regulatory proteins (G proteins). The G-protein subunits propagate signals to downstream effector molecules such as adenylyl cyclases, phospholipases, ion channels and MAP kinases [36].

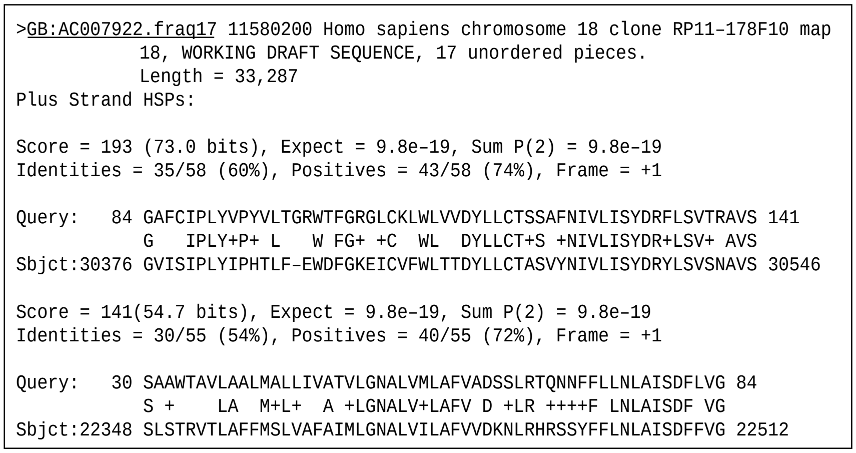
GPCRs have been divided into three main classes based on nucleotide and amino acid sequence similarity. These comprise family A (rhodopsin or adrenergic receptors), family B (calcitonin receptors) and family C (metabotropic receptors) [37]. The nature of the ligand can also be used as a basis for further classification. All GPCRs share a common overall structure in that they have seven hydrophobic TM α-helices (hence the commonly used term 7TM receptors), connected by three intracellular loops and three extracellular loops. The extracellular amino-terminal domains are of variable size, all with *N*-glycosylation sites, whereas the intracellular carboxy-terminal domains contain phosphorylation sites.

Recent examples of *in silico* identification of novel GPCRs

Selected publications illustrate the gene-finding approaches described previously. In 2001, a new histamine receptor (H4) was discovered independently by three different groups [38–40], using the human genome sequence as the data source. BLAST searches using the known histamine H3 receptor as a query sequence revealed an unordered, draft, contiguous genomic sequence encoding a putative partial novel GPCR, as shown in Fig. 2. After identification of the new receptor *in silico*, the first group used classical molecular cloning to isolate the full-length gene [38]. In the second publication [39], the entire coding region was predicted *in silico* using the Genewise program with an intron–exon structure identical to the known histamine H3 receptor [27]. The third group [40] used an alternative *in silico* approach based on the FAST-PAN program [41]. In this case, 20–50 protein sequences from a gene class were selected from the different branches of an evolutionary tree. These sequences were then searched against ESTs or genomic sequence databases using TFAST programs [42]. The sequence of this novel histamine receptor was confirmed by standard molecular biology techniques. The unique and discrete expression profile of H4 indicates a possible involvement in immune regulation, thus making it an attractive therapeutic target for pharmaceutical companies.

Another recent publication described the *in silico* identification of a novel family of GPCRs related to *Mas1* [43] called Mas-related genes (mrgs), which are expressed in sub-populations of sensory neurons [44]. In this example, a mouse GPCR was used to search both public and private (Celera, Rockville, MD, USA) genomic sequence databases using both BLAST and HMMs. Approximately 50 mouse genes were identified representing a family of highly duplicated genes. Several of these have been shown to be activated by RFamide peptides, a series of neuropeptides that have an amidated C-terminus and have arginine and phenylalanine as the final two amino acids. In contrast to the large number of receptors identified in the mouse, only eight intact human coding sequences belonging to the same family were identified, along with several pseudogenes. These new GPCRs might be therapeutic targets for pain because they are specifically expressed in certain subsets of nociceptive sensory neurons. However, the discrepancy in numbers between the two mammalian species raises interesting questions about the identification of true species orthologues for future target validation experiments. This contrasts with the older orphan GPCRs, GPR7 and GPR8, in which rodent orthologues exist for GPR7 but are absent for GPR8 [45]. Although neither of the human receptors was originally identified by the *in silico* methods described previously, bioinformatic searches for the rodent equivalent of GPR8 have undoubtedly been performed as reflected by the wealth of rodent sequence data now available, especially mouse sequence data. The absence of rodent orthologues poses a dilemma as generally in the pursuit of drug targets, part of the biological target validation process will involve rodent models. The same group that cloned GPR8 have recently reported a further 14 novel orphan GPCRs, all identified by computational searching of public domain high-throughput genomic sequence, patent and ESTs databases, with known mammalian GPCRs [46,47].

It is not just mammalian sequences that have been used to uncover novel human GPCRs. For example, homology searching of genomic sequence with a metabotropic glutamate receptor sequence from *Caenorhabditis elegans* ultimately led to the identification of two novel human family C GPCRs [48]. In this case, deciphering the open reading frame (ORF) in an identified genomic sequence was achieved with the Genscan gene prediction program. Verification of the predicted transcript by ESTs also led to the identification of a further closely related novel gene in the dbEST database that could be assembled into a complete ORF from an EST cluster.



```
>GB:AC007922.fraq17 11580200 Homo sapiens chromosome 18 clone RP11–178F10 map
           18, WORKING DRAFT SEQUENCE, 17 unordered pieces.
           Length = 33,287
Plus Strand HSPs:

Score = 193 (73.0 bits), Expect = 9.8e–19, Sum P(2) = 9.8e–19
Identities = 35/58 (60%), Positives = 43/58 (74%), Frame = +1

Query:   84 GAFCIPLYVPYVLTGRWTFGRGLCKLWLVVDYLLCTSSAFNIVLISYDRFLSVTRAVS 141
            G   IPLY+P+ L   W FG+ +C  WL  DYLLCT+S +NIVLISYDR+LSV+ AVS
Sbjct:30376 GVISIPLYIPHTLF–EWDFGKEICVFWLTTDYLLCTASVYNIVLISYDRYLSVSNAVS 30546

Score = 141(54.7 bits), Expect = 9.8e–19, Sum P(2) = 9.8e–19
Identities = 30/55 (54%), Positives = 40/55 (72%), Frame = +1

Query:   30 SAAWTAVLAALMALLIVATVLGNALVMLAFVADSSLRTQNNFFLLNLAISDFLVG 84
            S +    LA  M+L+  A +LGNALV+LAFV D +LR ++++F LNLAISDF VG
Sbjct:22348 SLSTRVTLAFFMSLVAFAIMLGNALVILAFVVDKNLRHRSSYFFLNLAISDFFVG 22512
```

DDT: InfoBiotech supplement

Figure 2. Discovery of the histamine H4 receptor [38]. The histamine H3 receptor protein (H3R) from nucleotide AF140538 was used to search the draft human genomic sequence using BLAST. A fragment of working draft sequence AC007922 was identified. The figure shows the amino acid alignment. 'Query' represents H3R, and 'Sbjct' is the new receptor. The alignment pairs roughly equate with exons. The regions of highest identity correspond with transmembrane domains (TMs), in this case, TMs 2 and 3. Further fragments of the genomic sequence AC007922 led to the identification of other parts of the receptor and enabled prediction of an *in silico*-derived partial cDNA. Full-length cDNA was identified in human bone marrow and peripheral blood mononuclear cells using molecular cloning techniques and sequence data derived from the predicted cDNA.

Identifying GPCR ligands *in silico*

Approximately 300 of the estimated 700 human GPCRs are non-sensory receptors, which are the most attractive drug targets. However, although novel GPCR identification is important, it comprises only one step in the process, and the key to understanding the function of GPCRs and harnessing their therapeutic potential, is to identify their ligands (endogenous or not). The ligands for GPCRs are structurally very diverse and range from very small peptides, such as neurotransmitter monoamines or amino acids, to large peptides or small proteins, such as chemokines. Less than half of the known non-sensory GPCRs are orphans (Steve Foord, pers. comm.). Unfortunately, it is not easy to predict peptide ligands computationally and most of those that have been discovered to date have been identified using classical biochemical techniques. The difficulties exist because, in most cases, they are relatively short (3–200 amino acids) and are processed from longer precursors (70–600 amino acids). Precursors contain either one neuropeptide [such as corticotropin-releasing factor (CRF)], multiple copies of the same neuropeptide [such as thyrotropin releasing hormone (TRH)], or copies of different neuropeptides [such as proopiomelanocortin (POMC)]. In addition, neuropeptides do not share a high level of amino acid similarity. All these factors thus make the identification of novel neuropeptides by simple homology searches difficult. However, there are some examples in which it has been possible to identify new peptides using this approach.

Two novel chemokines (CCL27/CTACK and CCL28) were identified by TBLASTN searches against private and public EST databases [49–51]. The identification of novel chemokines is easier because of their longer sequences and structural conservation. One human neuropeptide, cortistatin, was discovered by homology searching of human EST data sources, but this was simply the identification of the human orthologue of the experimentally identified rat preprocortistatin [52]. Although the entire human amino acid sequence has approximately 55% homology to the rat precursor, the predicted mature peptides are well-conserved. The most important difference is at the presumed dibasic amino acid cleavage sites, which leads to a 14-mer peptide in rat and a 17-mer peptide in humans. Both have similar biological activity.

Another approach is the use of HMMs. Urocortin II, a CRF neuropeptide, was identified computationally through this method [53]. HMMs were constructed using CRF family members from different species and the model was used to search the public domain human genome sequence. A genomic clone with significant homology was identified and one EST was used to extend the putative novel neuropeptide. Two additional CRF ligands were recently identified using motifs based on the primary and secondary structure typical of CRF peptides to search the GenBank genomic databases [54]. One of the ligands, SCP, was identified from a partial cDNA, and the other, SRP, from human genomic sequences. Motif searches, such as repetitive consensus motifs and upstream secretory signal peptide sequences, were used to identify additional RFamide-related peptides [55]. In this example, two ESTs and one genomic clone were identified. One novel human gene was characterized that encodes at least three novel RFamide peptides.

Alternative computational methods

Although these examples illustrate the possibility of using homology searches to identify *in silico* novel ligands, their identification remains difficult. Therefore, alternative computational methods are being developed. The use of inductive logic programming (ILP) [56] has been described to generate a grammar for neuropeptide precursors [57]. The grammar contains rules that describe neuropeptide precursors such as signal peptides. The best predictor makes the search for novel neuropeptide precursors more than a hundredfold more efficient than randomly selecting proteins from the SwissProt database.

Another example is the use of phylogenetic analysis by constructing trees to de-orphanize receptors. In this example, an efficient algorithm was developed and tested to associate ligands and receptors using a tree comparison [58]. To illustrate the potential of this alternative approach the authors tested the algorithm successfully with chemokine receptors.

As with direct computational identification of GPCRs, the discovery of ligands also requires the use of different approaches, as illustrated by the examples described here. For other therapeutically important target classes, similar *in silico* techniques can be applied with certain limited modifications.

Concluding remarks

Since the mid-1990s, most high-throughput biological data has been sequence data, first as ESTs and second as genomic sequence. Therefore, scientists in both industry and academia have been actively mining and analysing the information to identify novel targets. Although this research continues, the rate of new discoveries is declining. It is difficult to estimate how many targets identified *in silico* are now in the drug discovery pipeline of pharmaceutical companies; however, it is clear that some of the examples illustrated here, such as the identification of the H4 receptor, show promise. H4 is interesting not only because it is a therapeutic target, but also because it contributes to a better understanding of some of the pharmacological effects seen with other members of that sub-family of GPCRs.

The identification of a potential novel target by computational analysis of sequence data is only the first step in the long process of drug discovery. Bioinformatics methods and *in vitro* methods continuously interact along that pathway. We are now entering an era of high-throughput data-generation, and the contribution of sequence information is merely the first step in that revolution [59]. Other methods, such as microarrays and proteomics techniques, are now being routinely used for target identification and validation. These techniques are generating very large sets of functional data that need to be analysed and, perhaps more importantly, integrated. In fact, biological data integration is likely to be one of the biggest challenges facing informatics development during the first years of this decade. Although large datasets will be computerized, it has been recently recognized that a substantial amount of biological knowledge is still hidden in the literature. By automated analysis of MEDLINE records, networks of genes for gene expression have been created [60]. In fact, the term 'bibliome' has even been coined to describe this approach [61]. Once again, this is a classical example of how multiple computational approaches will be required to find useful information for the discovery of new drugs.

References

1 Lander, E.S. *et al.* (2001) Initial sequencing and analysis of the human genome. *Nature* 409, 860–921

2 Venter, J.C. *et al.* (2001) The sequence of the human genome. *Science* 291, 1304–1351

3 Deloukas, P. *et al.* (2001) The DNA sequence and comparative analysis of human chromosome 20. *Nature* 414, 865–871

4 Das, M. *et al.* (2001) Assessment of the total number of human transcripts units. *Genomics* 77, 71–78

5 Hogenesch, J.B. *et al.* (2001) A comparison of the Celera and Ensembl predicted gene sets reveals little overlap in novel genes. *Cell* 106, 413–415

6 Drews, J. (1997) Genomic sciences and the medicine of tomorrow. In *Human Disease – From Genetics Causes to Biochemical Effects* (Drews, J. and Ryser, S., eds.), pp. 5–9, Blackwell

7 Drews, J. and Ryser, S. (1997) The role of innovation in drug development. *Nat. Biotechnol.* 15, 1318–1319

8 Drews, J. (2000) Drug discovery: a historical perspective. *Science* 287, 1960–1964

9 Adams, M.D. *et al.* (1993) Rapid cDNA sequencing (expressed sequence tags) from a directionally cloned human infant brain cDNA library. *Nat. Genet.* 4, 373–380

10 Drake, F.H. *et al.* (1996) Cathepsin K, but not cathepsins B, L or S is abundantly expressed in human osteoclasts. *J. Biol. Chem.* 271, 12511–12516

11 Sakurai, T. *et al.* (1998) Orexins and orexin receptors: a family of hypothalamic neuropeptides and G-protein coupled receptors that regulate feeding behaviour. *Cell* 92, 573–585

12 Haseltine, W.A. (2001) Genomics and drug discovery. *J. Am. Acad. Dermatol.* 45, 473–475

13 Sanseau, P. (2001) Impact of the human genome sequencing for *in silico* target discovery. *Drug Discov. Today* 6, 316–322

14 Benson, D.A. *et al.* (2002) GenBank. *Nucleic Acids Res.* 30, 17–20

15 Horn, F. *et al.* (1998) GPCRDB: an information system for G protein-coupled receptors. *Nucleic Acids Res.* 26, 275–279

16 Hodges, P.E. *et al.* (2002) Annotating the human proteome: the human

proteome survey database (HumanPSD™) and an in-depth target database for G protein-coupled receptors (GPCR-PD™) from Incyte Genomics. *Nucleic Acids Res*. 30, 137–141

17 Le Novere, N. and Changeux, J-P. (2001) LGICdb: the ligand-gated ion channel database. *Nucleic Acids Res*. 29, 294–295

18 Rawlings, N.D. *et al*. (2002) MEROPS: the protease database. *Nucleic Acids Res*. 30, 343–346

19 Duarte, J. *et al.* (2002) NUREBASE: database of nuclear hormone receptors. *Nucleic Acids Res*. 30, 364–368

20 Smith, C.L. *et al*. (1997) The protein kinase resource. *Trends Biochem. Sci.* 22, 444–446

21 Rubin, G.M. *et al*. (2000) Comparative genomics of the eukaryotes. *Science* 287, 2204–2015

22 Bargmann, I. (1998) Neurobiology of the *Caenorhabditis elegans* genome. *Science* 282, 2028–2033

23 Enmark, E. and Gustafsson, J-A. (2001) Comparing nuclear receptors in worms, flies and humans. *Trends Pharmacol. Sci*. 22, 611–615

24 Nadler, M.J.S. *et al.* (2001) LTRPC7 is a Mg-ATP-regulated divalent cation channel required for cell viability. *Nature* 411, 590–595

25 Altschul, S.F. *et al.* (1990) Basic alignment search tool. *J. Mol. Biol*. 215, 403–410

26 Gelfand, M.S. *et al.* (1996) Gene recognition via splices alignment. *Proc. Natl. Acad. Sci. U. S. A.* 93, 9061–9066

27 Birney, E. and Durbin, R. (1997) Dynamite: a flexible code generating language for dynamic programming methods used in sequence comparison. *ISMB* 5, 56–64

28 Burge, C.B. and Karlin, S. (1997) Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol*. 268, 78–94

29 Uberbacher, E.C. and Mural, R.J. (1991) Locating protein-coding regions in DNA sequences by multiple sensor-neural approach. *Proc. Natl. Acad. Sci. U. S. A.* 8, 11261–11265

30 Altschul, S.F. *et al*. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*. 25, 3389–3402

31 Krogh, A. *et al*. (1994) Hidden Markov models in computational biology. Applications to protein modeling. *J. Mol. Biol*. 235, 1501–1531

32 Gough, J. and Chotia, C. (2002) Superfamily: HMMs representing all proteins of known structure. SCOP sequence searches, alignments and genome assignments. *Nucleic Acids Res*. 30, 268–272

33 Kim, J. *et al.* (2000) Identification of novel multi-transmembrane proteins from genomic databases using quasi-periodic structural properties. *Bioinformatics* 16, 767–775

34 Gill, R.W. *et al*. (1997) A new dynamic tool to perform assembly of expressed sequence tags, ESTs. *Comput Appl. Biosci.* 13, 453–457

35 Hubbard, T. *et al*. (2002) The Ensembl genome database project. *Nucleic Acids Res*. 30, 38–41

36 Marinissen, M.J. and Gutkind, J.S. (2001) G-protein-coupled receptors and signaling networks: emerging paradigms. *Trends Pharmacol. Sci.* 22, 368–376

37 Probst, W.C. *et al*. (1992) Sequence alignment of the G-protein coupled receptor superfamily. *DNA Cell Biol.* 11, 1–20

38 Zhu, Y. *et al*. (2001) Cloning, expression, and pharmacological characterisation of a novel human histamine receptor. *Mol. Pharmacol.* 59, 434–441

39 Liu, C. *et al*. (2001) Cloning and pharmacological characterisation of a fourth histamine receptor (H4) expressed in bone marrow. *Mol. Pharmacol.* 59, 420–426

40 Nguyen, T. *et al*. (2001) Discovery of a novel member of the histamine receptor family. *Mol. Pharmacol.* 59, 427–433

41 Retief, J.D. *et al*. (1999) Panning for genes – A visual strategy for identifying novel gene orthologs and paralogs. *Genome Res.* 9, 373–382

42 Pearson, W.R. (1994) Using the FASTA program to search protein and DNA sequence databases. *Meths. In Mol. Biol*. 25, 365–389

43 Young, D. *et al.* (1986) Isolation and characterization of a new cellular oncogene encoding a protein with multiple transmembrane domains. *Cell* 45, 711–719

44 Dong, X. *et al.* (2001) A diverse family of GPCRs expressed in specific subsets of nociceptive sensory neurons. *Cell* 106, 619–632

45 Lee, D.K *et al.* (1999) Two related G protein-coupled receptors: the distribution of GPR7 in rat brain and the absence of GPR8 in rodents. *Mol. Brain Res.* 71, 96–103

46 Lee, D.K. *et al*. (2001) Identification of four novel human G protein-coupled receptors expressed in the brain. *Mol. Brain Res.* 86, 13–22

47 Lee, D.K. *et al.* (2001) Discovery and mapping of ten novel G protein-coupled receptors. *Gene* 275, 83–91

48 Robbins, M.J. *et al.* (2000) Molecular cloning and characterisation of two novel retinoic acid-inducible orphan G-protein-coupled receptors (GPRC5B and GPRC5C). *Genomics* 67, 8–18

49 Morales, J. *et al.* (1999) CTAK, a skin-associated chemokine that preferentially attracts skin-homing memory T cells. *Proc. Natl. Acad. Sci. U. S. A.* 96, 14470–14475

50 Wang, W. *et al.* (2000) Identification of a novel chemokine (CCL28), which binds CCR10 (GPR12). *J. Biol. Chem*. 275, 22313–22323

51 Pan, J. *et al*. (2000) A novel chemokine ligand for CCR10 and CCR3 expressed by epithelial cells in mucosal tissues. *J. Immunol.*165, 2943–2949

52 Fukusumi, S. *et al.* (1997) Identification and characterization of a novel human cortistatin-like peptide. *Biochem. Biophys. Res. Commun.* 232, 157–163

53 Reyes, T.M. *et al.* (2001) Urocortin II: a member of the corticotropin-releasing factor (CRF) neuropeptide family that is selectively bound by type 2 CRF receptors. *Proc. Natl. Acad. Sci. U. S. A.* 98, 2843–2848

54 Yu Hsu, S. and Hsueh, A.J.W. (2001) Human stresscopin and stresscopin-related peptide are selective ligands for the type 2 corticotropin-releasing hormone receptor. *Nat. Med.* 7, 605–611

55 Hinuma, S. *et al.* (2000) New neuropeptides containing carboxy-terminal RFamide and their receptor in mammals. *Nat. Cell Biol.* 2, 703–708

56 Muggleton, S. and Raedt, L.D. (1994) Inductive logic programming: theory and methods. *J. Logic Programming* 19, 629–679

57 Muggleton, S.H. *et al.* (2001) Are grammatical representations useful for learning from biological sequence data? A case study. *J. Comput. Biol*. 8, 493–521

58 Bafna, V. *et al*. (2000) Ligand-receptor pairing via tree comparison. *J. Comput. Biol*. 7, 59–70

59 Butler, D. (2001) Data, data everywhere. *Nature* 414, 840–841

60 Jenssen, T-K. *et al*. (2001) A literature network of human genes for high-throughput analysis of gene expression. *Nat. Genet*. 28, 21–28

61 Searls, D.B. (2001) Mining the bibliome. *The Pharmacogenomics Journal* 1, 88–89