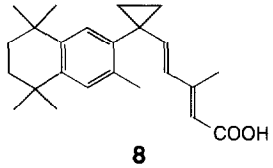
Retinoid X receptor ligands

As described previously [*Drug Discovery Today* (1997) 2, 501–502] retinoid hormones regulate many biological processes, such as cell differentiation, proliferation and apoptosis, through regulation of gene transcription. There are two families of retinoid receptors, the retinoic acid receptors (RAR) and the retinoid X receptors (RXR), each having three distinctive subtypes (α, β and γ). Although several retinoids have been recently studied in clinical trials for potential applications in oncology, there are high incidences of undesirable side effects. This toxicity has been associated with the ability of these retinoids to activate multiple retinoid receptors in various tissues. Several research groups are therefore attempting to develop more-selective retinoid-based agents with greater specificity of action. The group from Ligand Pharmaceuticals (San Diego, CA, USA), which was highlighted in *Monitor* last month, has described the design and synthesis of another series of potent RXR selective ligands [Farmer, L.J. *et al. Bioorg. Med. Chem. Lett.* (1997) 7, 2747–2752].

The most potent of these compounds (**8**) was shown to have minimal RAR agonist activity with good RXR binding (K_d = 4–5 nM) and transactivation (EC_{50} = 5–13 nM).



8

Genomics

Why ESTs are devolving: the 'full-length' story

Recent efforts within the pharmaceutical and biotechnology communities to complete human cDNA sequences have been spurred on by three factors. First, the need for functional information analogous to that coming from the genomes of yeast (*Saccharomyces cerevisiae*), flies (*Drosophila*) and worms (*Caenorhabditis elegans*) that links a gene to a pathology. Second, the requirement for structural information to satisfy molecular modelling efforts – for example, the ability to predict the nature of the small organic molecules that can fit into the active site of an enzyme to block its activity. Third, the need for full-length sequences to satisfy patent requirements and gain eventual priority battles. There are several cases where only one or a few misplaced bases differentiate published sequences, and these are likely to set a precedent for companies such as Myriad (Salt Lake City, UT, USA) and Incyte (Palo Alto, CA, USA) [Thomas, S.M. and Burke, J.F. *Expert Opin. Ther. Pat.* (1997) 7, 565–569].

The comparison of homologous cDNAs is 90% of the battle in the identification of the corresponding protein function from a novel sequence. The more complete the sequence, the better the chance of finding a homologous cDNA. Not only does an accurate, intact sequence allow reliable homology searches for related sequences, but it also permits a comparison of protein secondary structures and, consequently, functions [Lima, C.D., Klein, M.G. and Hendrickson, W.A. *Science* (1997) 278, 286–290]. The present algorithms accept 30% nucleotide sequence homology, even when the 3D structures of the gene products deviate, if the whole cDNA sequence is known. The availability of full-length sequences allows comparison of 3D structures where there is little or no sequence homology [Grausz, J.D. *Drug Discovery Today* (1997) 2, 510–511, with reference to the work of Dr Rajeev Aurora at Johns Hopkins University, Baltimore]. Thus, without a full-length cDNA sequence, there is the risk of false positive identifications, and also of missing potentially related genes through false negatives. It is extremely important to emphasize the need to use the amino acid sequence and the longest open reading frame (ORF) in a putative messenger, especially when the sequence is novel and cannot be found on the public databases. In this sense, expressed sequence tags (ESTs) are no longer a valid end point of research, but rather act as a start point for finding the full-length sequence of a gene.

Sequencing 'novel' genes

The concept of 'novel' genes deserves some discussion, because these genes are both the most difficult to analyse and the most intriguing with respect to physiology and pathology. When exploring novel cDNAs it is easy to charge up blind alleys, especially when the sequence data are poor and the correct ORF is not obvious. First, there are splice alternatives, so it is extremely important to work almost exclusively with a single, nonchimeric clone (i.e. full-length libraries make sense). Second, there are genes with no obvious ORF, either because they are expressed as an active RNA (e.g. XIST) or they have very complex patterns of alternative splicing (e.g. the TAT sequence). Third, relying upon genomic sequencing to fill the gap in a message can be very frustrating. The genome is a labyrinth, with regions of mirror-image sequences, sequence repetitions and noncontiguous jumps (i.e. transcripts that incorporate information from noncontiguous regions within or between chromosomes). Also, genomic sequencing may involve fruitless sequencing of a long intron or a quasi-identical region. [There are at least three documented cases of the latter (e.g. near SMA), where there is an adjacent tract of direct or mirror-image sequence repetition to mislead the uninitiated genomic traveller.] Overall, the depth of analysis possible is a function of the clones that have been sequenced (see Table 1).

While Table 1 demonstrates the clear analytical advantages of preparing full-length cDNA clones and sequencing candidate single clones from such libraries, actual methods of preparing full-length cDNA libraries are not always obvious. Protocols vary as a function of the availability/frequency, stability and primary sequence (e.g. presence of CpG stretches, palindromes and sequence repetitions) of the mRNA and the number of splice alternatives and/or close relatives in the same tissue/cell line. The published procedure involves direct

  1359-6446/98/$19.00. PII: S1359-6446(97)01139-2

cloning of selected, long first strands of cDNA by dC tailing, followed by poly-dG priming of the second strand, restricting the ends to leave four-base overhangs, and then ligating into an expression plasmid of choice [Defer, N. *et al. FEBS Lett.* (1994) 351, 109–113]. Recent improvements in reverse transcriptase, the availability of thermoresistant enzymes, and topoisomerase ('fidelase') have simplified the procedure, but the better methods are proprietary and guarantee reliability only in the hands of experienced investigators.

Sequencing full-length inserts to obtain full-length cDNAs

There is a significant technical leap from completely sequencing the insert of a cDNA clone and ensuring that the sequence obtained goes from the 5′ end of the first exon to the 3′ end of the last exon. As the insert sequences are completed at an ever increasing pace, the need to 'cluster' these to ensure the integrity of their extremities is becoming more and more evident for several reasons.

First, the sequencing itself must be accurate. The most common errors in sequencing introduce apocryphal stop codons. As pointed out by Bernard Dujon, with reference to yeast sequences, the difference between 99.99% and 99.90% accuracy is that over 50% of the ORFs are lost [*Trends Genet.* (1996) 7, 263–270]. The ESTs most readily available are often only 95% accurate; thus, it is only when these are assembled into overlapping 'contigs' that the sequence is verified. The pseudo-stop codons leave our computers stuttering, making translation into amino acids difficult and structural comparisons impossible.

Second, the clones should be close to full-length representatives of the expressed portions of the genes. Spliced alternatives have to be distinguished, and overlapping pieces of sequence have to be assembled or 'clustered' by computer. As Figure 1 shows, this can take the form of spreading from nodes of overlap or filling from the ends of clones.

Third, the extremities of a sequence, especially the 5′ end, have to be completed. A common nested PCR-based method of assuring the integrity of the 5′ end is shown in Figure 2. In simple terms, one searches for the clones with the longest insert between a known sequence at the 5′ end of the cDNA insert and the cloning vector. The sequence of this long insert can then be compared with genomic DNA sequences to verify its correct origins.

Sequence information – the first step

There is an ever increasing flood of genetic information, linking mutations to pathology. However, it is often difficult or impossible to predict the best therapy to limit or even reverse the damage and, where possible, cure a patient. A coordinated effort centred around the *IMAGE* consortium (contact Charles Auffray, auffray@infobiogen.fr) and encouraged

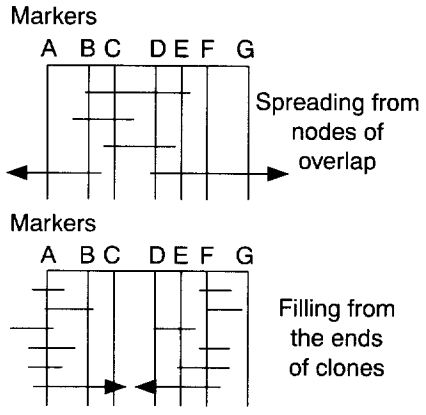


Figure 1. *Clustering of expressed sequence tags (ESTs) by computer can take the form of spreading from nodes of overlap or filling from the ends of clones.*

by the US Department of Energy (DOE), with both public and private participation, is intending to increase substantially (and ensure) the number of full-length cDNA clones long before the first human genome sequence is due to be completed in 2010. These efforts are being significantly encouraged by the *Merck Gene Index* (MGI). There are currently ~2,000 verified, complete human cDNA sequences in public databases, and maybe as many more in private databases. A specially designed database is being proposed by Patricia Rodriguez-Tome (tome@ebi.ac.uk) for the consortium to house the tens of thousands of sequences that will be available in the not too distant future.

Table 1. Possible analyses as a function of the available sequence information[a]

Analysis	ESTs	Clustered clones	Full length clones[b]	Full length + 5′ adjacent sequences[b]	Chimeric genomic clones	Intact genomic clones
Map relative to markers	+++	+++	+++	+++	++?	+++
Search sequence homology	+/−	++	+++	+++	+++	+++
Determine longest ORFs	−−−	+/−	+++	+++	++−	+++
Identify domains (functional, structural)	−−−	+/−−−	+++	+++	++−	+++
Determine 3D structure	−−−	+/−−	+++	+++	++?	+++
Determine function	−−−	+/−−	++/−	+++	++?	+++
Determine developmental program	−−−	−−−	+/−	++/−	++?	+++

[a]The level of ability to use data is qualitatively expressed as a combination of + and − signs, where + = likely, − = unlikely. The ? denotes that the data may or may not be sufficient to use. ESTs, expressed sequence tags; ORFs, open reading frames.

[b]Hybridization on arrays and differential display methods should yield considerable information on tissue localization and the place of a novel transcript/gene in the developmental program.

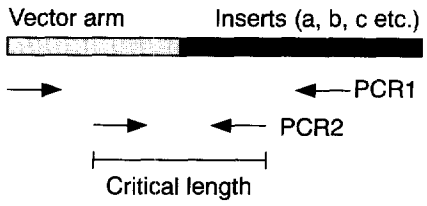


Figure 2. *A nested polymerase chain reaction (PCR)-based method for determining the 5′ end of cDNAs. The PCR product from the first reaction (PCR1) is isolated and then subjected to a second PCR reaction (PCR2) to increase the specificity. The PCR products from several clones (a, b, c, etc.) are isolated and the longest products are sequenced to give the complete 5′ end. The critical length defines the maximum length consistent with its genomic sequence.*

Cancer – an excellent example

Attempts to address the oncogenes and anti-oncogenes associated with particular tumours have been hindered by the basic genetic instability of neoplastic cells. With full-length sequence information, it is possible to identify the mechanism(s) responsible for instability and model the observed mutations into yeast, fly and worm systems. Scientists have begun both to classify tumours according to their drug susceptibility (reflection of the mutant proteins) and to design therapies. Several groups [Hartwell, L.H. *et al. Science* (1997) 278, 1064–1068; Weinstein, J.N. *et al. Science* (1997) 275, 343–349] have been classifying tumours and tumour cell lines by their drug responses. These studies have defined constellations of tumours as a function of such properties as *p53* alterations and cell-cycle abnormalities. Furthermore, they can predict the mutated genes by the pattern of drugs to which an unknown cancer (or cell line) shows heightened sensitivity.

A new potential-target-selection strategy, described by Hartwell and coworkers, using so-called synthetic lethal mutations, has resulted from the availability not only of complete expressed sequences but also of a complete genome sequence for yeast. A synthetic lethal mutation is one that compromises a gene in a parallel or a related pathway, rendering cells bearing the primary mutation (and only such cells) inviable. The secondary lethal mutation can be found by testing candidate loci, based upon a knowledge of biochemical pathways. Alternatively, a genome-wide scan of secondary sites can be performed to identify lethal mutations specific to yeast, fly or worm cells with the primary defect. Once such a lethal mutation is observed and the corresponding gene identified, a treatment can, theoretically, be developed in five further steps:

- Selection of pathways and alternative targets in yeast, flies or worms by introducing a homologous (to a putative human tumour locus) mutation and searching for a synthetic lethal mutation.
- Identification of the homologue to the alternative target, indicated by the synthetic lethal mutation, in mammals.
- Verification that the analogous mutations are synthetic lethal in mammalian cells. This will require a delayed knockout of the second target to demonstrate effectiveness against a growing tumour.
- Simultaneous evaluation of the pharmacological feasibility of inactivating the second gene. Generally an enzyme with a well-defined substrate is the target of choice.
- Initiation of high-throughput screening. Putative ligands for the chosen (synthetic lethal) target can first be designed by molecular modelling.

Summary

Genomics is presently undergoing a switch in emphasis from sequencing and mapping to the analysis of gene function – from 'structure' to 'function' of the genome [Hieter, P. and Boguski, M. *Science* (1997) 278, 601–602]. This reflects a more general move in biology to use pathology as an indication of gene function, rather than identifying genes to better understand and classify pathology. Thus, we are moving from a technology that classifies data to one that mines that data for valuable insights and extending the information based upon the biological hypotheses that evolve. This move has been made possible by the completion of the genomic sequence of the budding yeast, *S. cerevisiae*, and those of about a dozen prokaryotes. Short of the complete sequence of the human genome we can gain tremendous insight from studying complete expressed sequences of human genes and comparing them with homonyms in other eukaryotes.

Acknowledgements

I thank R. Mark Adams (Alphagene, Woburn, MA, USA), Daniel Caput (Elf Biorecherche, Labège, France) and Marvin Stodolsky (US Department of Energy).

J. David Grausz
50 Avenue Mathurin Moreau
75019 Paris, France
fax: +33 144 05 1970
e-mail: jd.grausz@chu-stlouis.fr

High-throughput screening

Managing the HTS laboratory

The role of screening in drug discovery has undergone enormous change. Twenty years ago, an academic postdoctoral scientist entering the industry might have considered the acceptance of a one-day-per-week screening commitment to be a necessary compromise in order to be able to do 'real science' during the other four days of the week (without the necessity of writing proposals for research grants).

By contrast, the modern screening laboratory is at the apex of the drug discovery process. Biomolecular screening is recognized as a scientific discipline in its own right, situated at the point of convergence of the many diverse technologies and sciences involved in generating a 'hit' and converting it into a lead compound.

  1359-6446/98/$19.00. PII: S1359-6446(97)01133-1