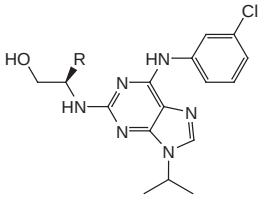
human tumour cells cultivated in hollow fibres. Three compounds have been identified that are currently being evaluated in human tumour xenografts.

Purine CDK inhibitors

Purines occur at relatively high concentrations in all living organisms where they play critical roles as cofactors and signalling molecules in modulating protein function. Furthermore, purines have been a fruitful source of cyclin-dependent kinase (CDK) inhibitors, targets that are especially attractive because of their key role in regulating the cell cycle. A recent publication describes the use of both solution- and solid-phase methods for the synthesis of purine-based libraries and the results of screening these compounds against CDK [Chang, Y-T. *et al.* (1999) *Chem. Biol.* 6, 361–375].

Supporting purine precursors on solid-phase allowed the combinatorial variation of two substituent positions. The effects of substituents in the 2-, 6- and 9-positions are additive and the results of screening the binary libraries against starfish oocyte CDK1–cyclin B enabled the identification of potent trisubstituted purine products. The most active compounds were subsequently tested for their ability to inhibit growth of U937 human leukaemia cells. In general, the cell-based IC_{50} values were higher than the *in vitro* values, presumably as a result of competition with high concentrations of intracellular ATP.



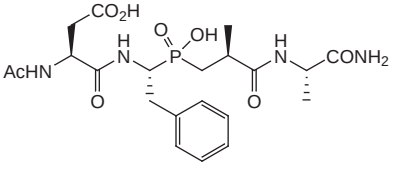
3 R = isopropyl
4 R = H

From the several hundred compounds prepared, four highly specific and potent CDK inhibitors were identified. These various purines act selectively on different biochemical pathways affecting cell-cycle progression. For example, compound (**3**) arrests the cell cycle specifically in the G2-phase whereas compound (**4**) induces M-phase arrest.

Selective ACE inhibitors

Inhibitors of angiotensin converting enzyme (ACE) have been employed for many years as effective treatments for hypertension, cardiac failure and diabetic nephropathy. These drugs block the renin-angiotensin cascade and prevent the formation of the hypertensive peptide, angiotensin II. Recently, it has been shown that the two catalytic domains of ACE might have slightly different functions. Evidence suggests that the *N*-domain might be responsible for the breakdown of peptides such as acetyl-seryl-aspartyl-lysyl-proline (AcSDKP), a negative regulator of haematopoietic stem cell differentiation and proliferation.

To help define the distinct roles of the two active sites, combinatorial chemistry has been used to identify compounds that selectively inhibit the *N*-domain site [Dive, V. *et al.* (1999) *Proc. Natl. Acad. Sci. U. S. A.* 96, 4330–4335]. A phosphinic peptide library has been used to identify a compound (**5**) that can differentiate the two ACE active sites



5

(*N*-domain site K_i = 12 nM, *C*-domain site K_i = 25 μM). Further studies with this compound might reveal the contribution of the ACE *N*-domain active site to the breakdown of AcSDKP.

Nick Terrett
Discovery Chemistry
Pfizer Central Research
Sandwich, Kent, UK
fax: +44 1304 655419
e-mail: nick_terrett@sandwich.pfizer.com

Bioinformatics

Rapid searching of sequence databases

Searching sequence databases is one of the most common tasks for any scientist with a newly discovered protein or nucleic acid sequence and is used to determine or infer:

- If the sequence has been found and already exits in a database
- The structure (secondary and tertiary)
- Its function or chemical mechanism
- The presence of an active site, ligand-binding site or reaction site
- Evolutionary relationships (homology).

Sequence database searching is different from database interrogation searching. Generally, sequence searching involves searching for a sequence in a database of sequences. By contrast, interrogation searching involves searching for keywords or other text in the text information (labelled the 'header') associated with each sequence in a database (SRS at EMBL http://www.emblheidelberg.de/srs/srsc is an example of a database interrogation program).

When comparing just two sequences, rigorous pairwise alignment algorithms can be used such as the Needleman and Wunsch algorithm [Needleman, S.B. and Wunsch, C. (1970) *J. Mol. Biol.* 48, 443–453] for optimal global alignments, and the Smith and Waterman algorithm [Smith, T.F. and Waterman, M.S. (1981) *J. Mol. Biol.* 147, 195–197] for optimal local alignments.

However, these two algorithms are computationally intensive and usually much too slow (with single central processing unit computer architecture) for searching a large database of sequences with a query (or unknown) sequence. Computers using parallel multi-processors can run these programs faster (see Box 1), but these computer architectures are expensive. Other algorithms, however, which employ heuristics (which essentially means they use some sort of assump-

 1359-6446/99/$ – see front matter PII: S1359-6446(99)01398-7

tion or mechanism that allows them to take a 'short cut' when comparing sequences) have been developed to speed up the process.

FASTA and BLAST

Until recently, the two most common heuristic database search algorithms were FASTA [Pearson, W.R. (1990) *Methods Enzymol.* 183, 63–98] and BLAST (basic local alignment search tool) [Altschul, S. *et al.* (1990) *J. Mol. Biol.* 215, 403–410]. There are a variety of related databases such as BLASTN, BLASTX, FASTX and TFASTX depending on the type of sequence used as the query and the type of database searched.

The principal heuristic of these algorithms is to use 'words' to search the database. A 'word' can consist of any multiple and arrangement of characters in the protein or nucleic acid alphabet. For example, a protein 'word' could be EL (a two-letter word), ELS (a three-letter word), or ELSE (a four-letter word). A word is also known as a k-tuple or w-tuple, which is essentially a derivation of the word 'multiple'.

The main assumption in a word-based method assumes that related sequences are more likely to share several common words. In most word-based methods, increasing the word size enables searches to be performed faster, but reduces the sensitivity of the search (larger words might miss a possible similarity). For reviews on how to use FASTA and BLAST in similarity searching, see Pearson, W.R. [*Protein Sci.* (1995) 4, 1145–1160].

The heuristic algorithms are not rigorous and there is a chance that a weak but significant similarity between two sequences will be missed. This is because sequences might have diverged to such an extent that no common words of a particular length remain. There is always a balance between sensitivity and selectivity [Pearson, W.R. (1990) *Methods Enzymol.* 183, 63–98]. Sensitivity is the ability to identify distantly related sequences. Increasing sensitivity usually increases the number of matches that are observed, but makes the search slower and increases background levels. By contrast, selectivity is the ability to avoid false positives (i.e. unrelated sequences with a high similarity score).

Both FASTA and BLAST essentially use a two-step process that first matches words in the query sequence to the same words in sequences stored in the database. They then use the word match to establish or 'seed' an alignment of one sequence with another. A score is calculated for the similarity of this alignment and either accepted or rejected as a possible match according to the chosen parameters. These two steps in the heuristic method are concomitant and fast when comparing one sequence against many others in a database. However, when comparing several sequences against each other (for example, when comparing whole genomes, large gene fragments or one sequence database with another), both FASTA and BLAST are much slower. The problem is that, in a large search, there are many more dissimilar alignments than similar alignments but the programs still calculate a score for all alignments established by the word 'seed'. This problem has been resolved recently by the development of new algorithms called RAPID, PHAT and SPLAT [Miller, C. *et al.* (1999) *Bioinformatics* 15, 111–121].

Box 1. Sequence database search programs on the Internet

Rigorous algorithms

Parallel multi-processors for rigorous database searching:
Bioccelerator at EMBL – http://shag.embl-heidelberg.de:8000/
Helix Systems at NIH – http://helix.nih.gov/

Heuristic algorithms

BLAST – http://www.ncbi.nlm.nih.gov/BLAST/
FASTA – http://fasta.bioch.virginia.edu/
RAPID – http://bionf.man.ac.uk/RAPID

RAPID, PHAT and SPLAT

The approach that Miller, C. and coworkers use is still a heuristic algorithm based on words, but with a difference. In particular, their approach separates the word-matching step from the sequence-alignment step that occurs concomitantly in both FASTA and BLAST. Separating these two steps is the key to the speed of this approach, because time is not wasted calculating alignments for many dissimilar sequences in the database. This approach is achieved by the algorithm RAPID (rapid analysis of pre-indexed data structures) that combines the word-matching step with an estimate of similarity without performing the alignment step. The estimate of sequence similarity is based upon the probability of word frequencies within a particular database. Thus, if a particular word is found frequently in both sets of sequences, and this is more frequent than would be expected by chance, then the similarity of the sequences can be scored according to this probability.

The crucial aspect of this approach is to determine the distribution of word frequencies upon which the estimate of probability is made. The authors recognize problems of sequence redundancy (which has the effect of making certain words appear more frequent) and database bias (such as the high A+T content in the yeast genome) that can affect the calculation of word probabilities. However, the empirical estimates of word probabilities from the EMBL (DNA)

database appears to be sufficient to provide a sensitive and selective algorithm for sequence similarity.

The PHAT (probabilistic Hough alignment tool) and SPLAT (smart probabilistic local alignment tool) alignment algorithms take the analyses further and are used to provide more detail than that provided by RAPID. So how fast is this new RAPID approach? Well, the authors used both RAPID and BLAST 2.05 to search the est10.dat database for vector contamination by comparing it with the vector.ig database. They found that RAPID (with PHAT to generate the ungapped alignments) performed the search in about 33 minutes while BLAST took approximately 493 minutes.

Steve Bottomley
Adjunct Research Fellow
Curtin University of Technology
Scientific Consultant, SciCon Pty Ltd
PO Box 1714, Subiaco
WA 6904, Australia
tel: +61 500 555 064
fax: +61 500 555 067
e-mail: ibottoml@info.curtin.edu.au

About *Monitor*...

Monitor provides an insight into the latest developments in the fields allied to drug discovery through brief synopses of recent publications and presentations together with expert commentaries on the latest technologies. There are two sections:
Molecules summarizes the chemistry, pharmacological significance and biological relevance of new molecules reported in the literature and on the conference scene.
Profiles offers commentary on promising lines of research, new technologies, emerging molecular targets, novel strategies and legislative issues.
We welcome topical contributions for inclusion as Profiles. Articles of approximately 500–1000 words in length should provide an accurate summary of the essential facts together with an expert commentary to provide a perspective. Authors should be aware that articles for publication in Monitor are subject to peer-review, and occasionally Monitor articles might be rejected or, as is more likely, authors could be asked to revise their contribution. Articles might also be edited after acceptance. Proposals for articles should be directed to: Dr Andrew W. Lloyd, Monitor Editor, Department of Pharmacy, University of Brighton, Moulsecoomb, Brighton, UK BN2 4GJ. tel: +44 1273 642049, fax: +44 1273 679333, e-mail: A.W.Lloyd@brighton.ac.uk