3 Shrimpton, C.N. *et al.* (2000) Development and characterization of novel potent and stable inhibitors of endopeptidase EC 3.4.24.15. *Biochem. J.* 345, 351–356

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

Bioinformatics

Multiple alignment – an essential bioinformatic tool

Multiple alignment is an important and essential tool for bioinformatics research[1,2]. Multiple alignments are alignments of three or more homologous or similar sequences (DNA, protein or RNA). Table 1 shows that, typically, the alignments are represented as a two-dimensional (2-D) table, where rows represent individual sequences and columns represent residue (or base) positions.

Multiple alignments are important because they extract 'meaning' from a mass of primary sequence data. In particular, they are a concise summary of sequence relationships among homologous sequences, and they provide information on the relationship between sequences to a particular gene (protein) family, they help find weak similarities (from distantly related proteins) using only sequence data, and they generate a representative sample for all the members of the sequence family.

The 'meaning' extracted from the primary sequence data is summarized in the consensus of multiple alignments. Table 1 shows the most common representation of a consensus as a single line (or a 'pseudo sequence'), which is added at the bottom of an alignment and consists of letters that show the alignment of conserved residues (or bases) within each column. The consensus can also be represented in other ways, including Hidden Markov Models, Profiles, PRINTS, BLOCKS, regular expressions and rules. For an excellent introduction to multiple alignment and bioinformatics in general, see Attwood and Parry-Smith[3].

Table 1. Multiple alignment as a 2-D table

	Residue (or base) positions				
	1	**2**	**3**	**4**	**5**
Sequence 1	D	R	I	V	G
Sequence 2	D	R	V	V	G
Sequence 3	D	G	L	V	A
Sequence 4	D	R	L	V	G
Sequence 5	D	G	A	V	A
Consensus	D	r	–	V	g

A capital letter is a fully conserved residue, a lower case letter is a partially conserved residue.

The essential concepts of multiple alignments are that:

- There are families of sequences that share some common feature(s) – in structure or function or both – and this is reflected by the presence of conserved regions in the multiple alignment of those sequences.
- More homologous sequences in an alignment improves the chance that the sequence variation observed between them represents the variation that exists in the whole family of related proteins.
- The alignment of a family of protein sequences provides more information than the alignment of any pair of those sequences. That is, when three or more sequences are aligned, there is information in the combined sequences that is not present in any one sequence or any pair of sequences.

If information is required on the level of the relationship between sequences, or between sequences and an established sequence family, then multiple alignments must be comprised of homologous sequences. Multiple alignments can be made from similar (non-homologous) sequences to identify regions of local similarity, but then cannot be used to infer relationships between the sequences.

The many uses of multiple alignments include:

- Protein 3-D modelling
- Secondary structure prediction
- Molecular evolution – phylogeny
- Source of secondary databases (e.g. Blocks and Prints)
- Database searching
- Design of primers for PCR, drugs and vaccines
- Consensus calling in DNA sequencing
- Models for evaluation of crucial residues in enzyme action and ligand binding.

Multiple alignment methods

Figure 1 shows the general procedure for gathering the homologous sequences required for a multiple alignment. Figure 2 shows the general procedure after a multiple alignment has been generated and Figure 3 shows the two general

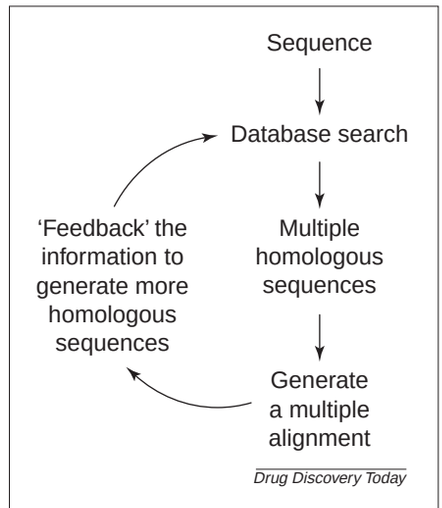

Figure 1. *Procedure for generating a multiple alignment.*

 1359-6446/00/$ – see front matter PII: S1359-6446(00)01504-X

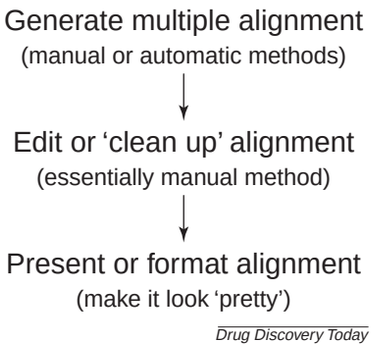

Figure 2. *Editing and formatting a multiple alignment.*

methods for generating a multiple alignment. As with pairwise alignment so there are both global and local multiple alignments. Global multiple alignment is useful for small sequences of similar length (usually <500 residues) and that share a global similarity. Local multiple alignment is useful for long sequences of unequal length (usually greater than 500 residues) that share isolated regions of similarity (usually separated by variable-length segments of little or no similarity). Alignment methods can be either automatic or manual.

Automatic methods use computer programs based on a particular theoretical approach and algorithm.

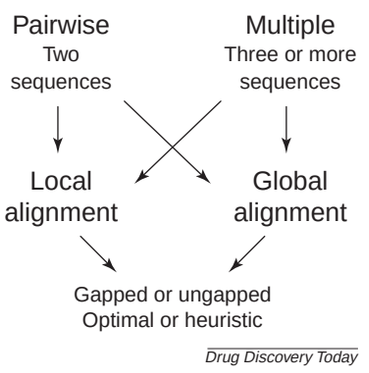

Figure 3. *Types of multiple alignment.*

Box 1 shows some examples of automatic multiple alignment programs. Automatic methods are used more often because multiple alignment of many sequences is very difficult and labour intensive. However, manual alignment – using experience and biological knowledge – is important and used to judge the efficacy of automatic methods and correct any obvious mistakes. Fortunately, interactive multiple alignment editors such as CINEMA (Ref. 4) make manual alignment a much more enjoyable experience.

Issues with multiple alignment

Any multiple alignment is not a 'perfect' alignment or even the most 'optimal' alignment of all the sequences. A multiple alignment is often an approximation based on the underlying theory, the particular algorithm, and the parameters used for a particular approach. For example, the ClustalX program[5] for multiple alignment aligns residues over the full length of the sequences comprising the alignment (it uses a progressive algorithm to produce a global multiple alignment). The alignment, in this case, depends on parameters such as the gap opening penalty, gap extension penalty, and the scoring matrix used to score the alignments. Different parameters will produce different alignments.

Furthermore, the rules for a consensus can be determined arbitrarily or based on weights calculated from a similarity or other scoring measure. In

Box 1. Examples of multiple alignment programs

Automatic methods

Clustal	http://www.csc.fi/molbio/progs/clustalw/
Dialign 2	http://bibiserv.techfak.uni-bielefeld.de/dialign/
MACAW	Can be downloaded from http://iubio.bio.indiana.edu/
Matchbox	http://www.fundp.ac.be/sciences/biologie/bms/matchbox_submit.html
Multalin	http://pbil.ibcp.fr/cgi–bin/npsa_automat.pl?page=/NPSA/npsa_server.html
MSA	http://www.ibc.wustl.edu/ibc/msa.html
Pileup	Found at various servers that implement GCG programs e.g. http://eBioinformatics.com/
PIMA	http://dot.imgen.bcm.tmc.edu:9331/multi-align/multi-align.html

Presentation methods

AMAS	http://barton.ebi.ac.uk/servers/amas_server.html
Consensus	http://www.bork.embl-heidelberg.de:8080/Alignment/consensus.html
MacBoxshade	Can be downloaded from http://iubio.bio.indiana.edu/
Pretty	Found at various servers that implement GCG programs e.g. http://eBioinformatics.com/
PrettyPlot	Found at various servers that implement GCG programs e.g. http://eBioinformatics.com/

Manual Editing

CINEMA	http://www.biochem.ucl.ac.uk/bsm/dbbrowser/CINEMA2.1/
SeqPup v(0.9)	Can be downloaded from http://iubio.bio.indiana.edu/soft/molbio/seqpup/java/

Note that the common multiple alignment programs, such as Clustal, can be found on many servers around the world and also have personal computer versions.

any event, the consensus is 'calculated' after a multiple alignment is performed and will vary depending on the assumptions or scoring methods used.

One way to help assess the validity and quality of the information in multiple alignments is to use at least two or three different methods (which involve different assumptions and underlying theory). If these methods produce similar results then it is likely that the results represent some real event.

1 Eddy, S.R. (1998) Multiple alignment & sequence searches. *Trends Guide to Bioinformatics* 15–17

2 Higgins, D.G. *et al.* (1996) Using CLUSTAL for multiple sequence alignments. *Methods Enzymol.* 266, 383–402

3 Attwood, T.K. and Parry-Smith, D.J. (1999) *Introduction to Bioinformatics*, Addison Wesley Longman Ltd (available at http://www.scicon.com.au).

4 Parry-Smith, D.J. *et al.* (1998) CINEMA – A novel colour interactive editor for multiple alignments. *Gene* 211, GC45–GC56

5 Thompson, J.D. *et al.* (1997) The ClustalX windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res.* 24, 4876–4882

Steve Bottomley
Curtin University of Technology
and SciCon Pty Ltd
Subiaco, Australia
e-mail: ibottoml@info.curtin.edu.au
URL: http//scicon.com.au

Would you like to contribute a review, news article, or meeting report to the *Drug Discovery Today* HTS supplement?
Have you any opinions on the articles we have published so far in this supplement series?
If so, please send your proposal or opinion to the Editor, Debbie Tranter, *Drug Discovery Today*, Elsevier Science London, 84 Theobald's Road, London, UK WC1X 8RR

Forthcoming articles in *the Drug Discovery Today* HTS supplement series
- High-throughput biology for functional genomics
- Early high-throughput pharmacokinetic and metabolic screening to aid selection of hits and leads
- Virtual screening and HTS
- Liquid handling for high-density plates
- Is there a future for robotics in HTS?
- Ultra-high-throughput techniques for SNP analysis
- High-throughput functional profiling of GPCR genome
- Assay design for HTS antibiotics drug discovery