KIND - a nonredundant database

Yvonne Kallberg & Bengt Persson

Department of Medical Biochemistry and Biophysics, Karolinska Institutet
Doktorsringen 9A3 5tr, S-171 77 Stockholm, Sweden
yvonne.kallberg@mbb.ki.se

Protein sequence databases are increasing in size at a rapid pace and offer a valuable source of information. In studies of biological variation, protein family characterisation and evolutionary relationships, these databases are of the outermost importance. However, as the databases expand, the time it takes to perform searches and extract the desired information increases as well.

Rapid search tools like Fasta and Blast is one way of dealing with this problem, another is to try to decrease the size of the databases without loosing valuable information. If we take a closer look at the content of databases we can see that the same sequence is present in several of them, i.e. the databases overlap. A protein sequence, or a variation of it might even appear in several entries within the same database. There are also subsequences, i.e. sequences that are part of a longer sequence, that could be removed without loss of information.

This redundancy aspect of protein sequence databases has been addressed several times, resulting in different solutions [e.g. 1]. We would like to present KIND (Karolinska Institutet Nonredundant Database), where Swissprot, Swissnew, Pir Trembl, GenPept and gpcu are merged into one, and identical sequences as well as subsequences are removed. A slightly modified version of the naive algorithm [2] is used for the matching process, where the sequences are matched in an all versus all manner. The database consists of around 300 000 entries, half of them originates from the protein databases and the other half from the translation of open reading frames (ORFs). Kind will be updated every 4-6 weeks, and is available via anonymous ftp from ftp://ftp.mbb.ki.se/pub/KIND.

References

1 Hermjakob, H., Lang, F. and Apweiler, R. (1998) SPTR - A comprehensive, non-redundant and up-to-date protein sequence database *The Bioinformer issue*

2 Gusfield, D. (1997) *Algorithms on strings, trees, and sequences: computer science and computational biology,* Cambridge University Press.