Nucleic Acids Research, 2001, Vol. 29, No. 1 **17–21**

The EMBL nucleotide sequence database

Guenter Stoesser*, Wendy Baker, Alexandra van den Broek, Evelyn Camon, Maria Garcia-Pastor, Carola Kanz, Tamara Kulikova, Vincent Lombard, Rodrigo Lopez, Helen Parkinson, Nicole Redaschi, Peter Sterk, Peter Stoehr and Mary Ann Tuli

EMBL Outstation-The European Bioinformatics Institute (EBI), Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK

Received September 18, 2000; Accepted September 27, 2000

ABSTRACT

The EMBL Nucleotide Sequence Database (http://www.ebi.ac.uk/embl/) is maintained at the European Bioinformatics Institute (EBI) in an international collaboration with the DNA Data Bank of Japan (DDBJ) and GenBank at the NCBI (USA). Data is exchanged amongst the collaborating databases on a daily basis. The major contributors to the EMBL database are individual authors and genome project groups. Webin is the preferred web-based submission system for individual submitters, whilst automatic procedures allow incorporation of sequence data from large-scale genome sequencing centres and from the European Patent Office (EPO). Database releases are produced quarterly. Network services allow free access to the most up-to-date data collection via ftp, email and World Wide Web interfaces. EBI's Sequence Retrieval System (SRS), a network browser for databanks in molecular biology, integrates and links the main nucleotide and protein databases plus many specialized databases. For sequence similarity searching a variety of tools (e.g. Blitz, Fasta, BLAST) are available which allow external users to compare their own sequences against the latest data in the EMBL Nucleotide Sequence Database and SWISS-PROT.

INTRODUCTION

The EMBL Nucleotide Sequence Database (http://www.ebi.ac.uk/embl/) represents Europe's primary collection of nucleotide sequences. The database is maintained at the European Bioinformatics Institute (EBI), an Outstation of the EMBL Molecular Biology Laboratory (EMBL) in Heidelberg, Germany. Data are received from genome sequencing centres, individual scientists, the European Patent Office (EPO) and via exchange from collaborating databases DDBJ (Japan) (1) and GenBank (USA) (2). To achieve optimal synchronisation, new and updated data are exchanged on a daily basis amongst the International Nucleotide Sequence Database Collaboration DDBJ/EMBL/GenBank. Users need only submit to one of the databases, irrespective of where the sequence will be published. The three databases adhere to a set of documented guidelines (The DDBJ/EMBL/GenBank Feature Table Definition) which regulate the content and syntax of the database entries. These guidelines ensure that the data continue to be made available in a format that can be exchanged efficiently between the databases, is compatible with current bioinformatics software and reflects developments in the fields of molecular and general biology. Established in 1980, the database was historically tightly coupled to the publication of sequences in the scientific literature. Electronic submissions via the WWW are now usual practice. Today, the vast majority of data submitted by direct transfer of data comes from major sequencing centres, such as the Sanger Centre. The EMBL database has nearly tripled in size within the last 11 months and on September 1, 2000 contained more than 9.6 Gigabases in 8.3 million records. Database statistics are available (http://www3.ebi.ac.uk/Services/DBStats/). A complete overview of the dataflow to and from the EMBL Database is provided in Figure 1.

DATA ACQUISITION

Direct submission systems

Most journals require authors to submit their sequence data to the sequence database as a prerequisite for journal publication. The EBI provides sequence submission systems including facilities for providing and checking biological information. A WWW-based interactive vector scanning service is available for submitters to assist in the screening of sequences for vector contamination before submission. The vector screening service uses the latest implementation of the BLAST algorithm and the special sequence databank EMVEC, comprised of a selection of sequences from the SYNthetic division of EMBL commonly used in cloning and sequencing experiments. EMVEC is updated with each release of EMBL and is available from the EBI's ftp server.

Webin. Webin is EMBL's interactive web-based system for submission of nucleotide sequences to the database. Webin is designed to allow fast submission of single, multiple or very large numbers of sequences**.** Webin collects all the information required to create a database entry: submitter information, release date information, sequence data, description and source information, reference citation information, feature information (e.g. coding regions, regulatory signals). WebIn is available at http://www.ebi.ac.uk/embl/submission/webin.html.

*To whom correspondence should be addressed. Tel: +44 1223 494466; Fax: +44 1223 494472; Email: stoesser@ebi.ac.uk

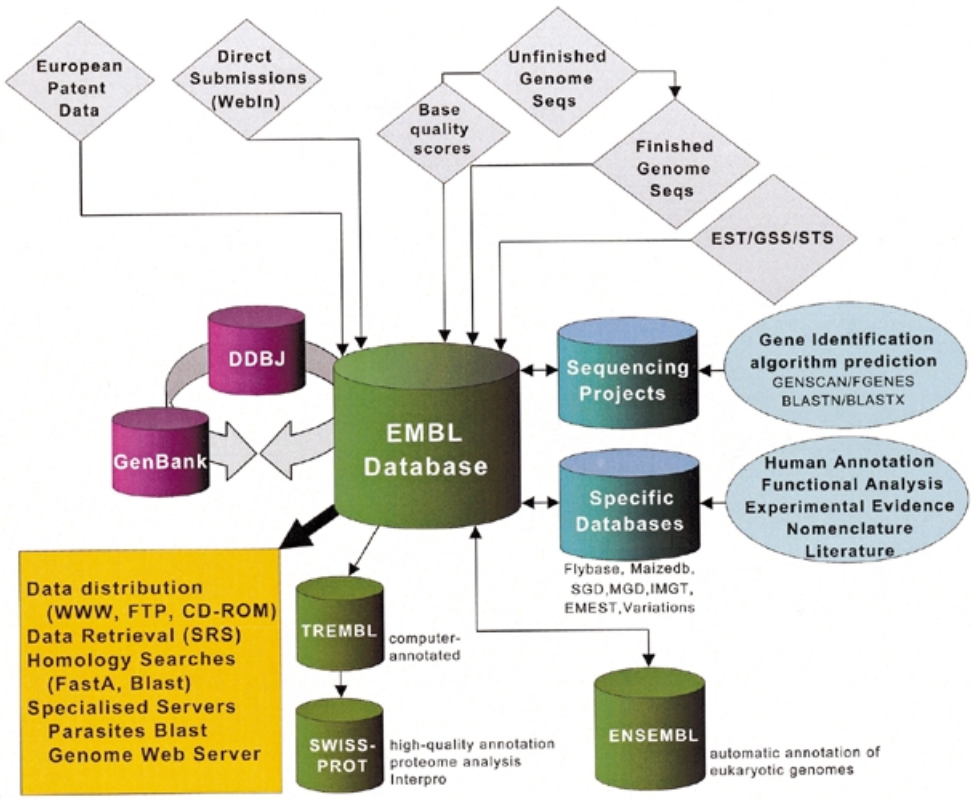

Figure 1. Dataflow EMBL nucleotide sequence database.

Sequin. Sequin is a stand-alone software tool developed by the NCBI for submitting and updating nucleotide sequences to the GenBank, EMBL or DDBJ databases. Sequin contains a number of built-in validation functions for enhanced quality assurance and runs on Macintosh, PC/Windows and UNIX computers.

Accession numbers. Accession numbers are unique identifiers that permanently identify sequences in the database. Accession numbers are assigned and communicated to authors within two working days of receipt of submission. These accession numbers (e.g. X64011 and AJ289709) are required by most biological journals before manuscripts are accepted. The suggested wording for citing a sequence in a publication is 'These sequence data have been submitted to the DDBJ/EMBL/GenBank databases under accession number AJ123456'.

Data confidentiality and release dates. During the submission process submitters specify whether their submitted data can be made available to the public immediately or whether the data should be withheld until an author-specified date. Data are never withheld after publication.

Sequence alignment submissions. Webin-Align is EMBL's interactive web-based system for submission of alignment data from phylogenetic and population analysis of nucleotide sequences. Unique alignment numbers (e.g. DS32096) are assigned to each alignment submission and should be included in the published article. Currently accepted standard alignment formats include NEXUS, PHYLIP, CLUSTAL and GCG/MSF or SEQUIN/ASN.1 output. Alignment data received at the EBI are curated by EMBL Database biologists and made available on EBI's network servers. As an additional service to the community, protein sequence alignments are also accepted and made available from the EBI FTP server. Nucleotide alignment data can be retrieved from the EBI's WWW pages at http://www3.ebi.ac.uk/Services/align/listali.html or from the FTP server at ftp://ftp.ebi.ac.uk/pub/databases/embl/align. Submission information is available from http://www.ebi.ac.uk/embl/Submission/

Updating existing database entries. Over time an entry which was correct when created may become out of date: authors may make corrections to the sequence itself, or may discover new features which require annotation. Since such findings are often not published, it is important that authors communicate their new findings to the database. The preferred option is via the WWW update form at URL: http://www3.ebi.ac.uk/Services/webin/update/update.html

Patent data

Patent Sequences are being captured in an ongoing collaboration with the EPO. EPO's policy is to release data to the public (and to EMBL) 18 months after the patent application date, regardless of whether a patent has been granted or not. Immediately after release by the EPO the latest patent sequence data are integrated into the EMBL database and made available to the public. All entries derived from the EPO patent literature are available from ftp://ftp.ebi.ac.uk/pub/databases/embl/patent/. Additionally, these files include data from American and Japanese patent literature incorporated from NCBI (USA) and DDBJ (Japan).

Genome Project data

Large-scale sequencing projects have become the major sources of new sequence data. The EMBL database opens submission accounts for groups producing large volumes of nucleotide sequence data over an extended period. Database entries produced at the research site are deposited and updated directly by the genome project submitter using FTP or email. For full details see http://www3.ebi.ac.uk/Services/GenomeSubm/. Groups wishing to make use of this submission procedure should contact the database at: datasubs@ebi.ac.uk. Sequence data will be included in the database as soon as they become available from the individual sequencing groups, and will immediately become available for homology searches via network services. High-throughput sequence records are included in the HTG division and contain keywords to indicate the finishing status of the sequencing (i.e., HTGS_PHASE0, HTGS_PHASE1, HTGS_PHASE2)

BIOLOGICAL ANNOTATION AND DATA CURATION

The importance of careful curation of individual sequences submitted directly by individual researchers and discussed in the scientific literature is obvious. Such sequences have often been the subject of experimental research elucidating features and function, while genome project submissions in most cases will 'only' include preliminary gene annotations based on gene prediction programs. Sequence annotation are an essential part of EMBL sequence records and current database policy is to reject submissions for which no sequence annotation has been provided, unless these describe ESTs or unfinished high throughput genome sequences (HTGs). In particular, it is essential to provide locations of coding regions, even when partial or preliminary, to allow inclusion of the corresponding translated protein sequence in the protein databases TrEMBL and SWISS-PROT (3). A team of biological curators review and check newly submitted data ensuring all mandatory information has been provided, that biological features are adequately described and that the conceptual translations of any coding regions obey universal translation rules. It has been necessary to automate many of the steps involved in checking new entries to cope with the overwhelming volume of new submissions. For this reason our submission tools now incorporate facilities for checking and providing additional taxonomic information and for ensuring that coding regions are correctly described. Curators may also suggest to submitters how the feature table may best be used to describe the biological features of particularly complicated or unusual data.

WWW guides

Internet guides to help submitters annotate their sequences are available from the EMBL–EBI WWW site and from within Webin.

- WebFeat. A complete list of feature table key and qualifier definitions, providing full explanations of their use.
- EMBL Annotation Examples. A selection of EMBL approved feature table annotations for some common biological sequences (e.g. ribosomal RNA, mitochondrial genome).
- DE Line Standards. Guidelines and database conventions for creating suitable descriptions for submissions.

DATA MANAGEMENT AND REPRESENTATION

Data is managed in a robust database management system (ORACLE), using a scheme which facilitates integration and interoperability with other databases, especially protein sequences. Quarterly releases and daily updates for distribution and installation at remote sites are generated from this system.

Database entry structure

Database entries are distributed in EMBL flat file format which is supported by most sequence analysis software packages and also provides a structure that is easy to read. The EMBL flat file comprises a series of strictly controlled line types (for details see User Manual at http://www.ebi.ac.uk/embl/Documentation/) presented in a tabular manner and consisting of four major blocks of data:

- Descriptions and identifiers. Entry name, molecule type, taxonomic division and total sequence length (found in the ID line); accession number (AC); sequence identifier and version (SV); date of creation and last update (DT); brief description of the sequence (DE); keywords (KW); taxonomic classification (OS, OC) and links to related database entries (DR).
- Citations. The citation details (RX, RA, RT and RL) of the associated publications and the name (RA) and contact details (RL) of the original submitter.
- Features. Detailed source information, biological features comprised of feature locations, feature qualifiers, etc.
- Sequence. Total sequence length, base composition (SQ) and sequence.

Sequence identifiers

In addition to unique and stable accession numbers, EMBL database entries include new sequence identifiers and versions that specify changes in sequences. The identifiers themselves remain stable within a given entry, whilst the version number increments with every sequence update. Protein identifiers can be used by external databases (such as SWISS-PROT) as an identifier onto which cross-references can be built at feature level, e.g. to individual CDS features. Protein identifiers are currently assigned to all protein translations of coding (CDS) features in the nucleotide sequence database to identify the exact protein translation for each coding sequence. These protein identifiers can be found in the Feature Table qualifier /protein_id.

Nucleotide sequence identifier. `Example: SV AJ400848.1`

Protein sequence identifier. `Example: /protein_id="CAB88705.1"`

Protein translations

Translations of protein coding regions represented by CDS features in EMBL entries are automatically added to the TrEMBL protein database. From these entries, SWISS-PROT curators subsequently create the SWISS-PROT database entries. EMBL nucleotide entries are cross-referenced (via the /db_xref qualifier) to the TrEMBL and SWISS-PROT databases.

Integration with other databases

Where appropriate, EMBL Database entries are cross-referenced to other databases like the Eukaryotic Promoter database (4) TRANSFAC (5), IMGT (6), Flybase (7), TrEMBL and SWISS-PROT. SWISS-PROT itself is linked to more than 30 different databases thus providing a focal point for database interconnectivity. Cross-references to external databases are represented in the EMBL flat file line type 'DR' and where appropriate, at the feature level via the feature qualifier /db_xref.

Database divisions

The EMBL Database currently consists of 18 divisions with each entry belonging in exactly one division. The division is indicated using three letter codes, e.g. PRO = Prokaryotes, HUM = Human, PHG = Bacteriophages, PLN = Plants, etc. The grouping is mainly based on taxonomy with a few exceptions like the HTG, EST, STS and GSS (Genome Survey Sequences) divisions. For these divisions grouping is based on the specific nature of the underlying data.

GENOME REPRESENTATION

Completed genomes web server

The first completed genomes from viruses, phages and organelles were deposited into the EMBL Database in the early 1980s. Since then, hundreds of complete genome sequences have been added to the database, including Archaea, Bacteria and Eukaryota. Direct access to completed genome sequences is available at http://www.ebi.ac.uk/genomes/

CON division

Among the database collaboration a new database division (CON) is being developed which will represent complete genomes, or other long sequences constructed from segmented entries. Each CON division entry will have an accession number and will contain information on how the construct is built from segments. In addition, the complete entry containing the full sequence, features and references will be retrievable through SRS.

High-throughput genome sequences (HTGs)

The HTG division includes 'unfinished' genome project data with annotation for many of the records being generated through computer analyses. This allows genome sequences produced by high-throughput sequencing projects to be available to the user community as soon as possible. The HTG division includes 'unfinished' genome project data with annotation for many of these records being generated through computer analyses. Entries in this division all contain keywords to indicate the status of the sequencing (e.g. HTGS_PHASE1). A single accession number is assigned to one clone, and as sequencing progresses and the entry passes from one phase to another, it will retain the same accession number. Once 'finished', HTG sequences are moved into the appropriate primary EMBL taxonomic division. EMBL Release 63 (June 2000) included over 3756 Mb of unfinished HTG data, compared to 544 Mb in Release 60 (September 1999).

Base quality values

Quality scores (Phrap) from draft HTG data are available on the EBI FTP server at ftp://ftp.ebi.ac.uk/pub/databases/embl/quality_scores. The gzip'ed files in the directory contain base quality values for unfinished human sequences from Japanese, US and European sequencing centres. The FastA-type headers contain the EMBL sequence identifiers and versions of the corresponding database entries. Quality score files are updated on a daily basis.

Draft human genome

A consortium of five publicly funded sequencing centres announced the completion of the human draft genome on June 26, 2000. Major sequence contributors consisted of the Sanger Centre, Cambridge, UK (http://www.sanger.ac.uk) and four American centres including Washington University Genome Sequencing Centre, St Louis (http://genome.wustl.edu/gsc/), Whitehead Institute for Biomedical Research, Cambridge, MA, (http://www-genome.wi.mit.edu/), Baylor College of Medicine, Houston, TX (http://www.hgsc.bcm.tmc.edu/) and DOE Joint Genome Institute, Walnut Creek, CA (http://www.jgi.doe.gov/).

Human draft sequence data can be accessed at the EBI via:

- Ensembl. Ensembl provides automatic annotation of the human draft genome data. Ensembl information includes confirmed peptides, confirmed cDNAs and predicted peptides. Repeat prediction along with integration of map information and SNPs are also available. Ensembl is a joint project between the Sanger Centre and EMBL–EBI and is accessible at http://www.Ensembl.org/. Additionally, a master web site ' Human Genome Central' contains a list of regularly updated links to useful human genome project resources thus providing a spring board for human genome data (http://www.Ensembl.org/genome/central/)
- Genome MOT. The Genome Monitoring Table (MOT) (8) presents the status of a number of large eukaryotic genome sequencing projects on the WWW. The tables are updated daily and provide access to individual EMBL database entries.
- EMBL release. Draft sequence data are also included in the EMBL Database HTG and HUM divisions: ftp://ftp.ebi.ac.uk/pub/databases/embl/release/.

DATA DISTRIBUTION, SEARCHING AND SEQUENCE ANALYSIS

EBI network services

Database releases are produced quarterly, and integrated into the EBIs SRS server. Databases and software can also be downloaded from the EBIs FTP server. EBIs network services allow access to the most up-to-date data collection via the Internet. Data access to EMBL nucleotide sequence data is also granted via email using the netserver or interactively via the WWW where the main service comprises the SRS server.

Sequence retrieval system (SRS)

The SRS server at the EBI integrates and links a comprehensive collection of specialised databanks along with the main nucleotide and protein databases. The SRS system (9) allows the databases to be searched using a number of fields including sequence annotations, keywords and author names. Complex querying and linking across all available databanks can also be executed and users should refer to the detailed instructions which are available online at http://srs.ebi.ac.uk/

Sequence searching

The EBI provides a comprehensive set of sequence similarity algorithms that can be accessed both interactively from the EMBL EBI WWW site (http://www.ebi.ac.uk/Tools/) or by email. The EMBL Nucleotide Sequence Database can be searched as a whole or by individual taxonomic division. The most commonly used algorithms available are Fasta3 (10) and WU-Blast2 (11; WU-blast HELP page). Fasta3 will find a single high-scoring gapped alignment between the query nucleotide sequence and database sequences. Comparisons between a nucleotide sequence and the protein databases can be made using fastx/y3, whilst tfastx/y3 allows comparisons between a protein sequence and the translated DNA databank. The EBIs Smith and Watermann (12) service comprises a comprehensive set of programs. These include today Compugen's Bic_SW, MPsrch (reference—see help page) and Scanps (reference—see help page). These facilitate more sensitive searches against protein sequence databases.

Sequence analysis

Specialised sequence analysis programs are also available from the EBI. Such services include multiple sequence alignment and inference of phylogenies using CLUSTALW (13), Gene prediction using GeneMark (14), pattern searching and discovery using PRATT (15), Motif identification using ppsearch (reference—see help page) as well as applications which have been developed in-house for various other projects.

EMBnet

The European Molecular Biology Network (http://www.embnet.org) was initiated in 1988 to link major European laboratories that provide bioinformatics to national scientific communities as well as being involved in active R&D in the fields of sequence analysis. One of the main tasks of the EMBnet network is the maintenance and updating of remote copies of the nucleotide and protein sequence databases which are updated daily. As bioinformatics grows, EMBnet plays an important role in providing a comprehensive program of bioinformatics training aimed specifically at both the wet lab researcher as well as programmers and systems administrators. A full listing of sites maintaining daily updated copies of the EMBL Database is available from the EBI at http://www.ebi.ac.uk/embl/Access/other_sites.html

CITING THE EMBL DATABASE

The preferred form for citation of the EMBL Nucleotide Sequence Database is: Stoesser,G., Baker,W., van den Broek,A.E., Camon,E., Garcia-Pastor,M., Kanz,C., Kulikova,T., Lombard,V., Lopez,R., Parkinson,H., Redaschi,N., Sterk,P., Stoehr,P. and Tuli,M.A. (2001) The EMBL Nucleotide Sequence Database. *Nucleic Acids Res.*, **29**, 17–21.

CONTACTING THE EMBL DATABASE

EMBL Nucleotide Sequence Submissions, European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK

Telephone:	
For data submissions	+44 1223 494499
General	+44 1223 494444
Fax:	
For data submission	+44 1223 494472
General	+44 1223 494468
Computer network:	
For data submission	datasubs@ebi.ac.uk
For other enquiries	datalib@ebi.ac.uk
For updates/publication notification	update@ebi.ac.uk

SUPPLEMENTARY MATERIAL

Table of relevant URL links available at NAR Online.

REFERENCES

1. Tateno,Y., Miyazaki,S., Ota,M., Sugawara,H. and Gojobori,T. (2000) DNA Data Bank of Japan (DDBJ) in collaboration with mass sequencing teams. *Nucleic Acids Res.*, **28**, 24–26.
2. Benson,D.A., Karsch-Mizrachi,I., Lipman,D.J., Ostell,J., Rapp,B.A. and Wheeler,D.L. (2000) GenBank. *Nucleic Acids Res.,* **28**, 15–18.
3. Bairoch,A. and Apweiler,R. (2000) The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res.,* **28**, 45–48.
4. Périer,R.C., Praz,V., Junier,T., Bonnard C. and Bucher P. (2000) The Eukaryotic Promoter Database (EPD). *Nucleic Acids Res.*, **28**, 302–303.
5. Wingender,E., Chen,X., Hehl,R., Karas,H., Liebich,I., Matys,V., Meinhardt,T., Prüß,M., Reuter,I. and Schacherer,F. (2000) TRANSFAC: an integrated system for gene expression regulation. *Nucleic Acids Res.*, **28**, 316–319. Updated article in this issue: *Nucleic Acids Res*. (2001), **29**, 281–283.
6. Ruiz,M., Giudicelli,V., Ginestoux,C., Stoehr,P., Robinson,J., Bodmer,J., Marsh,S.G.E., Bontrop,R., Lemaitre,M., Lefranc,G., Chaume,D. and Lefranc,M.-P. (2000) IMGT, the international ImMunoGeneTics database. *Nucleic Acids Res.,* **28**, 219–221. Updated article in this issue: *Nucleic Acids Res*. (2001), **29**, 207–209.
7. The FlyBase Consortium (1999) The FlyBase database of the Drosophila Genome Projects and community literature., *Nucleic Acids Res.* **27**, 85–88.
8. Beck,S. and Sterk,P. (1998) Genome-Scale DNA sequencing where are we? *Curr. Opin. Biotechnol.*, **9**, 116–121.
9. Etzold,T., Ulyanov,A. and Argos,P. (1996) SRS: information retrieval system for molecular biology data banks. *Methods Enzymol.*, **266**, 114–128.
10. Pearson,W.R. (1994) Using the FASTA program to search protein and DNA sequence databases. *Methods Mol. Biol*., **24**, 307–331.
11. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
12. Smith,R.F. and Waterman,M.S. (1981) Comparison of biosequences. *Adv. Applied Math*., **2**, 482–489.
13. Thompson,J.D., Higgins,D.G. and Gibson,T.J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, **22**, 4673–4680.
14. Borodovsky,M. and McIninch,J. (1993) GeneMark: Parallel Gene Recognition for both DNA Strands. *Comput. Chem.*, **17**, 123–133.
15. Jonassen,I., Collins,J.F. and Higgins,D.G. (1995) Finding flexible patterns in unaligned protein sequences. *Protein Sci.*, **4**, 1587–1595.