Quality control in databanks for molecular biology

CODATA Task Group on Biological Macromolecules and Colleagues[1]

Summary

Using a scientific measurement without an estimate of its error is like lending money to a stranger. Given the explosion in nucleic acid and protein sequence and structural data, what risks are the scientific and medical communities running in using these databases. Is there an 'ombudsman' who speaks for the users of the data? CODATA, the Committee on Data for Science and Technology of the International Council of Scientific Unions was established to improve the quality, reliability, processing, management, and accessibility of data for science and technology. The CODATA Task Group on Biological Macromolecules has surveyed quality control procedures of archival databanks in molecular biology. Our role is 'to advise, to be consulted, and to warn.' This report describes the kinds and extents of errors that may appear in nucleic acid and protein databases, and presents an agenda for future work to improve the quality of these databases. The results of the survey appear on the web (http://www.codata.org/codata/tgreports/tg_reps.html). *BioEssays* **22:1024–1034, 2000.**
© 2000 John Wiley & Sons, Inc.

Introduction

Databanks in molecular biology exist in profusion.[1] Some are general and comprehensive; others are specialized or "boutique" collections. The main archival projects curate, archive and distribute sequences and structures; these include the output of genome sequencing projects and the systematic whole-organism protein structure determinations, known as "structural genomics" (Table 1).

The quality of archived data can, of course, be no better than the data determined in the contributing laboratories.

[1]E. E. Abola, A. Bairoch, W. C. Barker, S. Beck, D. A. Benson, H. Berman, G. Cameron, C. Cantor, S. Doubet, T. J. P. Hubbard, T. A. Jones, G. J. Kleywegt, A. S. Kolaskar, A. Van Kuik, A. M. Lesk,* H.-W. Mewes, D. Neuhaus, F. Pfeiffer, L. F. TenEyck, R. J. Simpson, G. Stoesser, J. L. Sussman, Y. Tateno, A. Tsugita, E. L. Ulrich, J.F.G. Vliegenthart
Funding agency: The Wellcome Trust.
*Correspondence to: Arthur M. Lesk, Chair, CODATA Task Group on Biological Macromolecules, Wellcome Trust Centre for Molecular Mechanisms in Disease, Cambridge Institute for Medical Research, University of Cambridge Clinical School, Wellcome/MRC Building, Hills Road, Cambridge, CB2 2XY, U.K. E-mail: aml2@mrc-lmb.cam.ac.uk
Addresses of all authors appear on the Web (http://www.codata.org/codata/tgreports/tg_reps.html).

Nevertheless, careful curation of the data can help to identify errors. Disagreement between duplicate determinations is, as always, a clear warning of an error in one or the other. Similarly, results that disagree with established principles *may* contain errors. It is useful, for instance, to flag deviations from expected stereochemistry in protein structures, but such "outliers" are not necessarily wrong. Furthermore, different kinds of data can act as checks on each other; e.g., the identification of putative sequencing errors through comparison of protein structures.[2]

The state of the experimental art is the most important determinant of data quality. For instance, the acquisition of older data was limited by the earlier techniques. Amino acid sequences of proteins used to be determined by peptide sequencing but now are translated from DNA or RNA sequences. One effect of the data explosion, however, is that most data are new data, governed by current technology.

Quality control procedures provide the second level of protection. Indices of quality, even if they do not permit error correction, can help scientists avoid basing conclusions on questionable data. Yet, because the community clamours for instant access, several databanks release entries in an 'immature' state, and only subsequently pass them through checking procedures. In consequence, errors often enjoy a longer existence than they would if caught immediately, especially if they are disseminated to local copies at databases in which subsequent corrections are not made. That errors exist is incontrovertible. Readers with an appetite for horror stories may consult Ref. 3 (see also Refs. 4–7).

DNA sequence data and annotation

Current DNA sequencing technology should reduce sequence error rates to as low as one base in 10000.[5,8–12] In genome projects each base is sequenced, on average, between 6 and 10 times, generally including at least one reading from each strand. Inconsistencies are checked by experts, and if the conclusion is still uncertain, by additional experiments. As a typical protein in a prokaryote is encoded by a kilobase of DNA, an error in gene sequencing of one isolated wrong base in 10000 corresponds very roughly to one amino acid error in $\sim 10-15$ proteins for a prokaryotic genome. In the human genome, in contrast, the most gene-dense regions contain only about 1 gene per ~ 10000 bases, with the average estimated at 1 gene per 30000 bases. The corresponding error

Table 1. Main archival data projects in molecular biology. Nucleic acid sequences are treated by a collaboration of GenBank, the EMBL Data Library and the DNA Data Bank of Japan. A second triple partnership deals with amino acid sequences of proteins: The Protein Information Resource, the Munich Information Centre for Protein Sequences, and the International Protein Information Database in Japan. SWISS-PROT, a collaboration between the University of Geneva and the European Bioinformatics Institute, treats amino acid sequences of proteins. The Protein Data Bank, Nucleic Acid Database and the BioMagRes Data Bank collect three-dimensional structural data. A related organization, the Cambridge (U.K.) Crystallographic Data centre archives structures of small molecules.

Name of Databank and home URL	Type of data	Location
GenBank www.ncbi.nlm.nih.gov/	Nucleic acid sequences	National Library of Medicine, Washington, DC, USA
EMBL Data Library www.ebi.ac.uk/ebi_docs/embl_db/ebi/topembl.html	Nucleic acid sequences	European Bioinformatics Institute, Hinxton, UK
DNA Data Bank of Japan www.ddbj.nig.ac.jp/	Nucleic acid sequences	National Institute of Genetics, Mishima, Japan
Protein Identification Resource www.nbrf.georgetown.edu/pir/	Amino acid sequences	Georgetown University, Washington D.C.
Munich Information Center for Protein Sequences (MIPS) speedy.mips.biochem.mpg.de/	Amino acid sequences	Max-Planck-Institute für Biochemie, Martinsried, Germany
International Protein Information Database in Japan (JIPID)	Amino acid sequences	Science University of Tokyo, Noda, Japan
Swiss-Prot www.expasy.ch/sprot/	Amino acid sequences	Geneva, Switzerland and Hinxton, UK
Protein Data Bank www.rcsb.org	Protein and nucleic acid structures	Research Collaboratory for Structural Bioinformatics, USA
Nucleic Acid Database ndbserver.rutgers.edu/	Nucleic acid structures	Rutgers University, New Jersey, USA
BioMagResBank www.bmrb.wisc.edu/	NMR structure determination	Madison, Wisconsin, USA
CarbBank www.ccrc.uga.edu	Primary structures of carbohydrates	Complex Carbohydrate Research Center, University of Georgia, USA

rate in translated amino acid sequences would be (very roughly) one amino acid substitution in 100 proteins. Larger-scale errors in sequence assembly, however, can also occur, especially in highly-repetitive regions. Missing a nucleotide can cause a frameshift error affecting the computed translation, making nonsense of an amino acid sequence.

In the short term, more efficient chemical reactions and better software should improve sequencing accuracy.[13–18] Algorithms can detect errors in coding sequences,[19] and their extension to non-coding sequences should be possible. Combined with proper annotation, they should considerably improve the assessment of sequence quality.

How do sequencing errors affect applications, such as database searching? Most algorithms seek inexact matchings, and cope with substitutions, deletions and insertions.[20–23] In most cases involving comparisons of sequences from different species, evolutionary divergence rather than error accounts for the differences between the probe sequence and the entries identified (Fig. 1). In database searches using

nucleic acid sequences simple deletions of one or two bases do not present problems. The translation to amino acid sequences of nucleotide sequences containing such errors, however, will involve frameshifts that garble the sequence disastrously and create very serious problems in database searching.

Although the archival DNA databases (EMBL/GenBank/DDBJ) carry out quality checks on every sequence submitted,[24] no general quality control algorithm is yet in widespread use.[25] To safeguard the target error rate of <1 in 10000 for the human genome project, the National Institutes of Health have carried out a cross-genome centre sequence checking exercise in the U.S.A., to be followed by an international exercise.

Annotation of nucleotide sequence data
Entries in nucleic acid sequence databanks contain, in addition to the DNA or RNA sequence itself, annotations that contain information about: the origin of the sequence; the
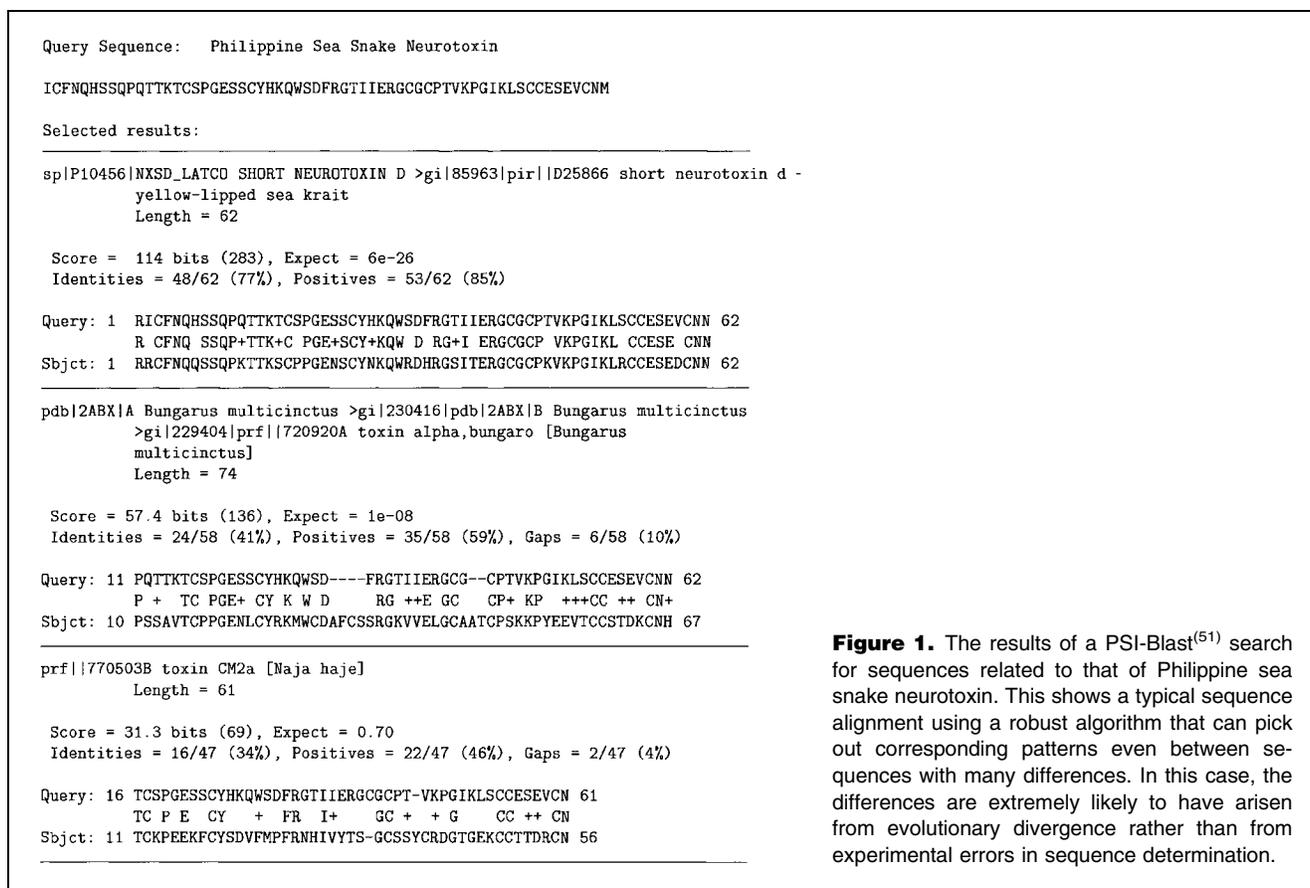
Challenges

```
Query Sequence:    Philippine Sea Snake Neurotoxin

ICFNQHSSQPQTTKTCSPGESSCYHKQWSDFRGTIIERGCGCPTVKPGIKLSCCESEVCNM

Selected results:
────────────────────────────────────────────────────────────────────

sp|P10456|NXSD_LATCO SHORT NEUROTOXIN D >gi|85963|pir||D25866 short neurotoxin d -
          yellow-lipped sea krait
          Length = 62

 Score =  114 bits (283), Expect = 6e-26
 Identities = 48/62 (77%), Positives = 53/62 (85%)

Query: 1   RICFNQHSSQPQTTKTCSPGESSCYHKQWSDFRGTIIERGCGCPTVKPGIKLSCCESEVCNN 62
           R CFNQ SSQP+TTK+C PGE+SCY+KQW D RG+I ERGCGCP VKPGIKL CCESE CNN
Sbjct: 1   RRCFNQQSSQPKTTKSCPPGENSCYNKQWRDHRGSITERGCGCPKVKPGIKLRCCESEDCNN 62


pdb|2ABX|A Bungarus multicinctus >gi|230416|pdb|2ABX|B Bungarus multicinctus
          >gi|229404|prf||720920A toxin alpha,bungaro [Bungarus
          multicinctus]
          Length = 74

 Score = 57.4 bits (136), Expect = 1e-08
 Identities = 24/58 (41%), Positives = 35/58 (59%), Gaps = 6/58 (10%)

Query: 11  PQTTKTCSPGESSCYHKQWSD----FRGTIIERGCG--CPTVKPGIKLSCCESEVCNN 62
           P +  TC PGE+ CY K W D     RG ++E GC   CP+ KP  +++CC ++ CN+
Sbjct: 10  PSSAVTCPPGENLCYRKMWCDAFCSSRGKVVELGCAATCPSKKPYEEVTCCSTDKCNH 67
────────────────────────────────────────────────────────────────────

prf||770503B toxin CM2a [Naja haje]
          Length = 61

 Score = 31.3 bits (69), Expect = 0.70
 Identities = 16/47 (34%), Positives = 22/47 (46%), Gaps = 2/47 (4%)

Query: 16  TCSPGESSCYHKQWSDFRGTIIERGCGCPT-VKPGIKLSCCESEVCN 61
           TC P E  CY   + FR I+    GC + + G   CC ++ CN
Sbjct: 11  TCKPEEKFCYSDVFMPFRNHIVYTS-GCSSYCRDGTGEKCCTTDRCN 56
────────────────────────────────────────────────────────────────────
```

Figure 1. The results of a PSI-Blast[51] search for sequences related to that of Philippine sea snake neurotoxin. This shows a typical sequence alignment using a robust algorithm that can pick out corresponding patterns even between sequences with many differences. In this case, the differences are extremely likely to have arisen from evolutionary divergence rather than from experimental errors in sequence determination.

investigators; links to other databanks; and the 'feature table,' a list of segments of the sequence known or thought to have biological significance—for instance, regions that code for proteins. Some annotations are hypothetical because many descriptive features are inferences derived from the sequences. For instance, identification of a gene coding for a protein in a stretch of genomic DNA from the distribution of bases and absence of stop codons is a deduction, not an observation. Such inferences have error rates of their own.

The hypothetical nature of many annotations is a new development. Until recently, the typical DNA sequence entry consisted of a gene, often with surrounding sequences, produced by a research group investigating that gene and its products. Annotation of features was grounded in experimental data. In contrast, in full-genome projects, there is in most cases little or no experimental work confirming the expression and characterizing the products of genes. Computer programs produce annotation, which have been selected/edited by skilled curators before release of the data.

In general, annotation of bacterial genomes is more complete and accurate than that of eukaryotes.[26] Bacterial genes are easy to identify because the spaces between them are small and they are free of introns. The types of errors that tend to appear are entries with frameshift sequencing errors, which lead to truncation of predicted reading frames or even double errors leading to a mistranslated internal fragment. Small genes, indeed any small functionally important sequences, are likely to be missed, as they may fall below statistically significant limits. In higher organisms, identifying genes is harder and, in consequence, database annotation is more dubious. Experimental studies can improve the annotation of genomic regions but it is impossible to guarantee the identification of all features of interest. Alternative splicing patterns present a particular difficulty.

The full sequence of the nematode worm, *C. elegans*, is typical of the complex genomes determined. Annotation of its genes is much harder than was that of the previous most complex genome sequenced, yeast. The *C. elegans* genome has the advantage of consistency, however, as it was sequenced entirely by two labs which collaborated extensively. In contrast, the sequence of the human genome is being determined in many labs and its annotation varies from nothing, for certain regions, to gene predictions that are based on different methods and that reflect different thresholds of accepted significance.

Because the raw data on which gene predictions and other annotations are based is increasing very steeply, consistency checks will become more comprehensive and statistical tests more powerful. Therefore the annotation of DNA sequences must be frequently updated and not frozen. It is a challenge for databanks to find ways to link primary sequence data to new and updated annotations.

Prospects

The rate of acquisition of DNA sequence data has been increasing exponentially for years, and new technology and new genome projects will ensure that this increase continues well into the future. Multiple capillary sequencing instruments have recently been commercialized that promise to produce 500,000 bases of raw DNA sequence per day per instrument. Hundreds of such instruments will be in place within a year. (This sequencing capacity could process the equivalent of one human genome per month.) Array scanning mass spectrometers will, even initially, have a similar capacity.

Both of these approaches appear to be scalable to even higher sequencing rates. This will almost certainly mean that upstream (sample preparation) and downstream (data annotation and assembly into finished sequence) processes will become rate limiting, rather than the production of sequence data themselves. A rapid explosion of production of sequence-related data such as expression patterns and polymorphisms is also expected to occur. Centralized annotation will be impossible, and some sort of controllable annotation process will need to be developed and diffused across the entire biological community.

Gel-electrophoresis-based DNA sequencing has error rates of 1% or more in the raw data. Improving the accuracy of the finished product by collation of multiple determinations complicates the data handling. As sequencing processing tools mature to include confidence estimates, these can be factored into algorithms that use the DNA sequences. In contrast, mass spectrometric sequence data are likely to be almost error free, given the absence of electrophoresis-specific artifacts like compressions and because of the enormous redundancy inherent in the data themselves, in particular the fact that the mass of a peak usually reveals the base composition of the fragment. Nucleic acid mass spectral data are typically accurate to fractions of a Dalton (Fig. 2).[27] This is quite a change from the data that molecular biologists usually confront!

Three-dimensional analysis

X-ray crystal structure analysis

The experimental data in an X-ray crystal structure determination are the structure factor magnitudes, the absolute values of the Fourier coefficients of the electron density. The results are estimates of the positions and effective 'sizes' of the atoms.

Contributions to the effective size of an atom include its vibrational amplitude and, more seriously for protein structures, the disorder in the region of the molecule that contains it. The reported parameter called the 'B-factor' of each atom describes its effective size, and for proteins it should be treated as an empirical value. Because every atom contributes to every observation, it is difficult to estimate errors in individual atomic positions.

The *resolution* of the data limits the potential quality of the structure. Resolution measures the ratio of the number of observations to the number of parameters to be determined. In the structure determination of small organic molecules or of minerals, this ratio is usually generous: ~ 10. But for a typical protein crystal, the following relationships hold:
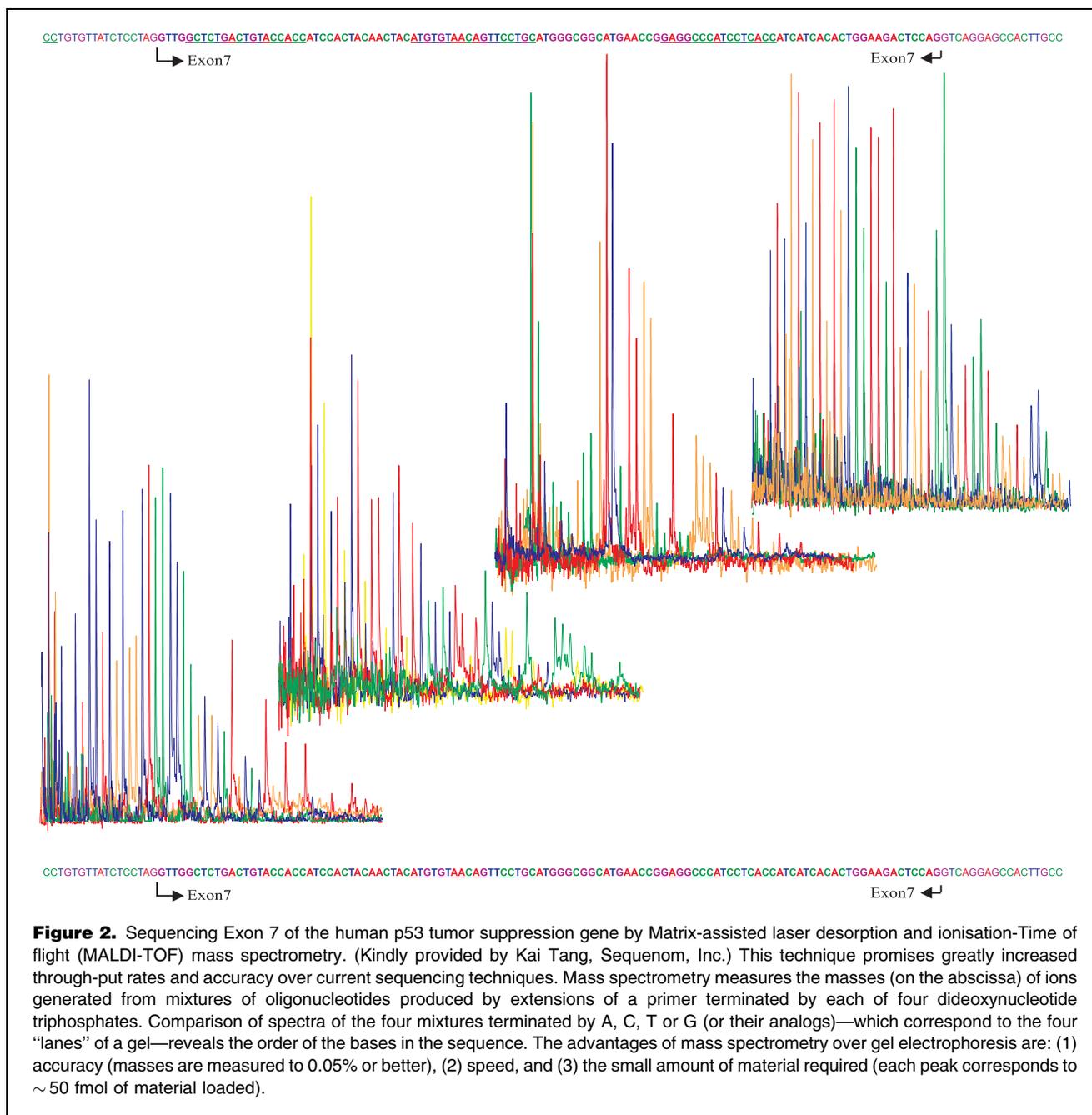
	Low resolution ... High					
Resolution in Å	4.0	3.5	3.0	2.5	2.0	1.5
Ratio of observations to parameters	0.3	0.4	0.6	1.1	2.2	3.8

(The median resolution of structures in the Protein Data Bank is about 2.0 Å.)

In practice, many models often fit the data almost equally well and crystal structure determinations of proteins usually require the imposition of stereochemical restraints such as standard bond lengths and angles. Furthermore, crystal structure determinations are at the mercy of degrees of order in different parts of the molecule. (Order is the extent to which different unit cells of the crystal are exact copies of one another.) An extreme case is immunoglobulin KOL, with data collected to very high resolution (1.9 Å) in which the variable domains were well determined but where the constant domains were completely disordered and invisible in the electron density map.[28] More typically, the core of a protein structure is well ordered but surface loops may be more mobile and less well determined.

The R-factor measures how well the model fits the data. If the set of observed X-ray intensities is Fo, and the corresponding predicted intensities calculated from the model are Fc, the R-factor is defined as $\Sigma|Fo - Fc|/\Sigma|Fo|$. (The set of F's may contain a list of tens of thousands of numbers.) For high-resolution models values around $0.18-0.22$ are good. For low-resolution studies, however, 'good' R-factor values may be obtained even for models that are largely or entirely wrong.[4,29,30] A more sophisticated quality measure is the cross-validation R factor, $R_{(free)}$[31] The definition of $R_{(free)}$ is the same as that of R except that the sum is taken over only a small subset (typically $\sim 5-10\%$) of the data, which have been withheld during the structure determination.

$R_{(free)}$ measures how well the model predicts data withheld during estimation of the parameters of the model. In an ideal world R and $R_{(free)}$ would be equal. In practice, when values of $R_{(free)}$ more than 0.1 greater than R are obtained, the result can

Figure 2. Sequencing Exon 7 of the human p53 tumor suppression gene by Matrix-assisted laser desorption and ionisation-Time of flight (MALDI-TOF) mass spectrometry. (Kindly provided by Kai Tang, Sequenom, Inc.) This technique promises greatly increased through-put rates and accuracy over current sequencing techniques. Mass spectrometry measures the masses (on the abscissa) of ions generated from mixtures of oligonucleotides produced by extensions of a primer terminated by each of four dideoxynucleotide triphosphates. Comparison of spectra of the four mixtures terminated by A, C, T or G (or their analogs)—which correspond to the four "lanes" of a gel—reveals the order of the bases in the sequence. The advantages of mass spectrometry over gel electrophoresis are: (1) accuracy (masses are measured to 0.05% or better), (2) speed, and (3) the small amount of material required (each peak corresponds to ~50 fmol of material loaded).

indicate significant problems with the structure. Other indicators of data quality are the resolution (the higher the better), the overall completeness of the data, the completeness of the highest resolution shell of data, and the average signal-to-noise ratio of the high-resolution data.

The final model in a structure determination is adjusted by a refinement procedure, and statistical analysis of the sensitivity to parameter variations of the fit of model to the experimental data can provide estimates of errors. Methods of error estimation, well-established for small molecules, have been extended to macromolecules.[32–33] Murshudov and Dodson[32], for example, estimate *overall* uncertainties of atomic positions in macromolecules from the $R_{(free)}$ values, giving in a typical case values of about ~0.05 Å at 1.5 Å resolution and ~0.15 Å at ~2 Å resolution. In addition, they approximate uncertainties of *individual* atomic positions from B-factors, giving values of about 0.16 Å for an atom with B = 20 Å,[2] and 0.3 Å for an atom with B = 60 Å.[2]
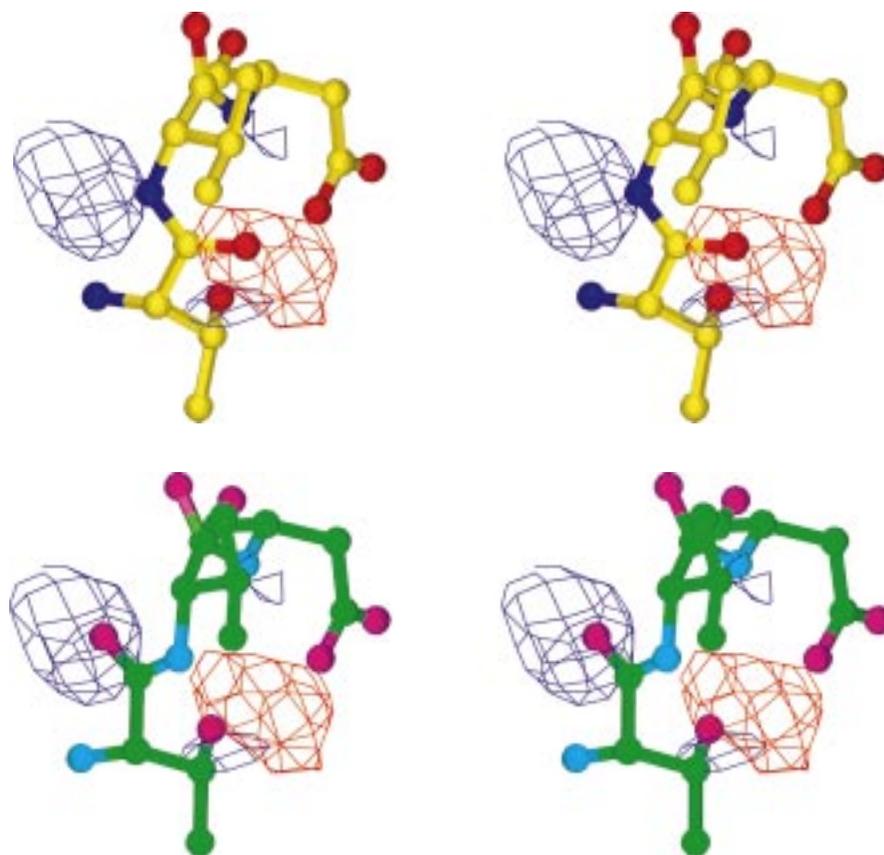
Figure 3. A difference map showing electron density and structural model superposed. Top: During the process of structure determination; bottom: corrected model. This figure illustrates the use of $(F)_{(0)}$-$(F)_{(c)}$ difference electron density maps. Shown are two residues (with yellow carbon atoms, blue nitrogen atoms, and red oxygen atoms), with—top illustration–the peptide link built incorrectly. The red contours show difference density at a negative level (i.e., a red peak means that there are atoms in the model which should not be there), whereas the blue contours show this density at a positive level (i.e., a blue peak means that there should be atoms there, but they are not in the model yet). In this case, the solution–shown in bottom illustration is to 'flip' the peptide plane, so that the carbonyl oxygen leaves the red density, and points into the blue density instead (model with green carbon atoms). The upper residue was also in an unfavourable region of the Ramachandran plot, but after "flipping" the peptide it ends up in a favourable area.

Errors in crystal structures reflect both errors in data and errors in solving the structure. Ohlendorf[34] compared four independently-refined structures of interleukin-1β determined at 2 Å resolution, finding a root-mean-square (r.m.s.) difference of 0.84 Å (a value much higher than the estimate obtained using the often-quoted method of Luzzati;[35] cf. remarks by Cruickshank Ref. 36). Fortunately, crystallographers are increasingly depositing their primary data, the structure factors, along with their structures. This permits detailed checks on the structures.[37]

In practical cases, scientists want a clear impression of the quality of a particular region of interest in a protein, such as an active site. They should examine the fit of the model to the data, by displaying the electron density map with the molecular model superposed on it.[38] The most useful maps for identifying problem areas in a structure are 'difference maps,' which have positive values where correct features are missing, and negative values where features of the model are not supported by the data (Fig. 3).

Flying blind: Assessment of quality of a set of protein coordinates without the experimental data from which it was derived

When experimental structure factors are available, assessment of a structure is a matter of checking the consistency of the results with the measurements. In many cases the structure factors are not available, however. How can one

Challenges

then check the structure? The B-factors are important clues; high B-factors, especially concentrated in a region, suggest that the region has not been well-determined. This usually reflects imperfect order in the crystal.

The other approach to structure validation in the absence of experimental data begins by flagging stereochemical "outliers"—exceptions to regularities common to well-determined protein structures. The difficulty is not in the *detection* of outliers but in deciding whether they are genuine features of the structure, or the result of errors in building an atomic model into the electron-density map, or the inevitable result of crystal disorder. Almost every macromolecule contains a few residues that will be flagged as outliers by validation software even though they may not be errors.

Several computer programs exist for finding outliers in structures: O, Procheck, Whatif, Errat, Verify3D.[39–41] Of course, the authors of these programs face the psychological/political problem of resentment that they are setting themselves up as "structure police." Scientists, like other people, do not welcome criticism, especially by individuals who did not contribute to the results and did not face the problems that arose during the work. Furthermore, given the expectation that the kudos for a good result should reflect the skill and effort that went into producing it, protein crystallography is a special field (as for analogous reasons is palaeontology), because crystallographers are at the mercy of their crystals (as palaeontologists are at the mercy of their fossil finds). Imagine a crystallographer investigating two proteins. One protein crystallizes in a well-ordered crystal, and data is collected to high resolution. The solution of the structure is virtually instantaneous and the result 'correct'; i.e. virtually free of outliers. The other protein forms poorly ordered crystals that diffract only to low resolution. There may be immense problems in interpreting the map, and the crystallographer, given a commitment to an interesting project, may sweat blood for years trying to extract the best possible results from the data. Yet even the best results achievable in such a case can only be of limited quality because of the resolution of the map. Of course, the structure police swoop down on the second structure.

Methods to detect outliers include:

- Type I: nomenclature and convention-related checks: Examples include incorrect chirality, and the naming of chemically equivalent side-chain atoms (e.g., in phenylalanine and tyrosine rings). Such errors can be corrected confidently without reference to experimental data and current submissions can be fixed at the time of deposition.[42] Checking of old datasets is in progress.[42]
- Type II: 'self-consistency tests:' Many stereochemical features of macromolecular models are restrained during refinement. Bond lengths and angles are restrained to ideal values, planarity is imposed on aromatic rings and carbo-

xylate groups, non-bonded atoms are prevented from clashing, temperature factors of atoms bonded to each other are forced to be similar, etc. Methods that assess how well these restraints are satisfied are an important part of the arsenal of structure verification tools. Nevertheless, their inadequacy in detecting genuine shortcomings in models has been demonstrated.[30]

Proper assessment of outliers (as features or errors) requires access to the experimental data. Sometimes, outliers warn of more serious problems and may require careful inspection of the electron-density maps and even model rebuilding by an experienced crystallographer. Unfortunately, not all errors can be fixed, even by appeal to structure factors and maps; some regions are fatally disordered.
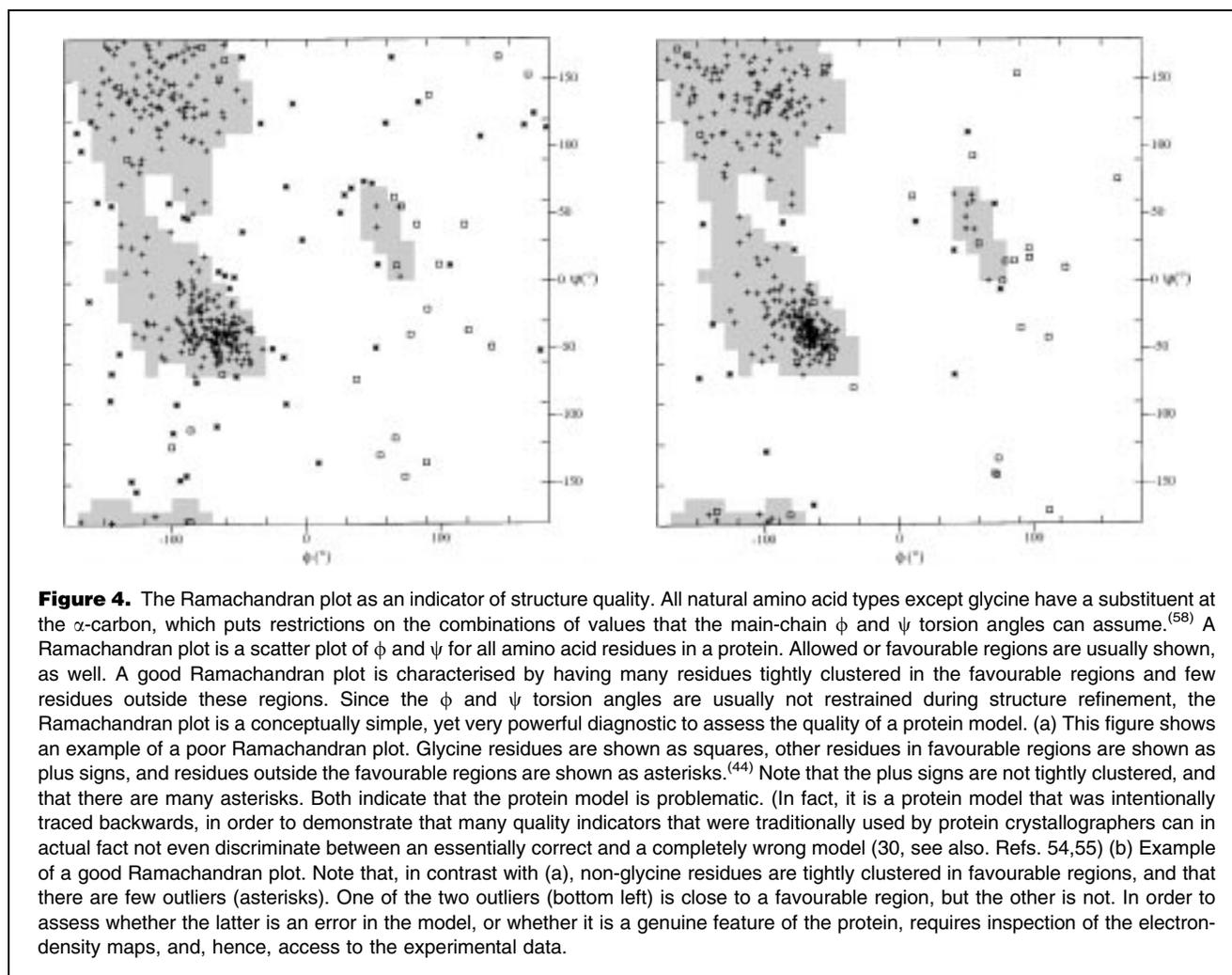
- Type III: 'orthogonal' tests: Most revealing and useful are verification methods independent of the restraints used during model refinement. Such methods use database-derived information to assess how usual or unusual an atom, residue, or entire molecule is. Examples include the analysis of torsion angles of the protein main-chain (Ramachandran analysis) and side-chain atoms (rotamer analysis), the orientation of the peptide plane (peptide-flip analysis), atomic volumes, geometry of the $C\alpha$-backbone, nonbonded contacts, and the use of sequence-structure profiles.[43]

For the non-expert user of macromolecular crystal structures, coordinate-based validation tools can help in forming a rough judgement regarding the quality of a model. In general, globally poor models (often determined at low-resolution) will give rise to very many outliers, particularly in the Type III tests. A similar heuristic is valid for *locally* poor models: unreliable parts of a model will be characterised by a concentration of 'unusual' residues, even though the global statistics may be acceptable. The simplest and most powerful test to carry out is inspection of the Ramachandran plot[44] (Fig. 4).

We emphasise that proper assessment of outliers requires access to the experimental data; and fixing of real errors will usually require the attention of an experienced crystallographer. The conclusion seems inescapable that structure factors should be archived and available, and we can think of no reason why a crystallographer who is willing to release the coordinates of a structure should hesitate to deposit the experimental data on which they were based.

Quality of NMR structure determination

NMR is the second major technique for determining macromolecular structure. The experiments determine approximate values of a set of interatomic distances and conformational angles. These distances, derived from the Nuclear Overhauser Effect (NOE), identify pairs of atoms close together in space, including those from residues distant in the sequ-

Figure 4. The Ramachandran plot as an indicator of structure quality. All natural amino acid types except glycine have a substituent at the α-carbon, which puts restrictions on the combinations of values that the main-chain ϕ and ψ torsion angles can assume.[58] A Ramachandran plot is a scatter plot of ϕ and ψ for all amino acid residues in a protein. Allowed or favourable regions are usually shown, as well. A good Ramachandran plot is characterised by having many residues tightly clustered in the favourable regions and few residues outside these regions. Since the ϕ and ψ torsion angles are usually not restrained during structure refinement, the Ramachandran plot is a conceptually simple, yet very powerful diagnostic to assess the quality of a protein model. (a) This figure shows an example of a poor Ramachandran plot. Glycine residues are shown as squares, other residues in favourable regions are shown as plus signs, and residues outside the favourable regions are shown as asterisks.[44] Note that the plus signs are not tightly clustered, and that there are many asterisks. Both indicate that the protein model is problematic. (In fact, it is a protein model that was intentionally traced backwards, in order to demonstrate that many quality indicators that were traditionally used by protein crystallographers can in actual fact not even discriminate between an essentially correct and a completely wrong model (30, see also. Refs. 54,55) (b) Example of a good Ramachandran plot. Note that, in contrast with (a), non-glycine residues are tightly clustered in favourable regions, and that there are few outliers (asterisks). One of the two outliers (bottom left) is close to a favourable region, but the other is not. In order to assess whether the latter is an error in the model, or whether it is a genuine feature of the protein, requires inspection of the electron-density maps, and, hence, access to the experimental data.

ence which are essential for assembling the overall folding pattern. Calculations then produce sets of structures that are consistent—as far as possible—with the experimental constraints on distances and angles, and that have proper stereochemistry.[45]

In practice, whereas crystallographers report unique (or only a small number of) structures, NMR spectroscopists usually produce a family of $\sim 10-20$ related structures (or even more) each calculated from a random starting point but all using the same set of experimentally-derived constraints (Figure 5). Comparison across such an ensemble is used to assess precision; regions in which the local variation in structure is small across the ensemble are considered well defined by the data. This may be viewed as the equivalent of the crystallographer's B-factor. For highly variable regions, the question arises whether this reflects genuine dynamic disorder or merely a lack of NMR data to fix this portion of the structure. This question can only be addressed by supplying additional

information, for instance by measuring NMR relaxation properties as a function of sequence in order to detect, experimentally, regions with significant internal mobility.[46] For many NMR structures, however, this information is unavailable.

In principle, the quality of fit between a calculated structure and the experimental data can be expressed using the NMR equivalent of the crystallographer's R-factor but this requires being able to *predict* NOEs from atomic coordinates. This is often difficult to do, in part because NOE intensities are strongly affected by internal motions. Hence, 'NMR R-factors' are not widely used for protein structures at present (but are more common for oligonucleotide structures). In practice, quality of fit to the experimental data is reported in terms of the maximum or average values for constraint violations. Conformity with stereochemical expectations is assessed as for X-ray structures, using measures such as average deviations from geometric ideality, proportion of residues occupying

Figure 5. Comparison of crystal and NMR structure determination of a small protein domain: the SH3 domain from α-spectrin. The crystal structure is in red. Four structures from the reported NMR ensemble of 15 models are black, blue, green, and magenta. They were selected to span the range of differences between NMR and crystal structures. The structures all agree well in the central core. There are small deviations in some of the peripheral loops. The agreement between crystal and NMR structures is relatively good, but not unusually so.

favoured regions of the Ramachandran plot, and calculated values of conformational energy.[47] Of course, in the calculations, the trade-off between constraint violation and geometric non-ideality can be adjusted by the user of the programs.

A key consideration is how effectively NMR data constrain a proposed structure. A common measure is the average number of constraints per residue. Constraints, however, differ in the extent to which they restrict conformation space—indeed, some are completely ineffective. Therefore the correlation of "constraints per residue" with structural quality is imperfect. By far the most important consideration is that constraints should all be correct, which in turn depends absolutely upon correct assignment of the spectrum, that is, the correct association of individual NMR signals with the corresponding atoms in the structure. It is this step of spectral assignment that probably represents the greatest difference between NMR and crystallographic structure determinations, and which affords the greatest opportunity for disasters. Fortunately, however, these are rare.

A decision that strongly affects apparent quality of NMR structure determinations is the choice of how many structures to report. If structural statistics are calculated using only the best few from a large number of calculated structures, the outcome will appear more attractive than if a larger ensemble is used. There is no consensus on how many structures to report, partly because calculation protocols differ widely in their convergence rates. Some control over this aspect of reporting is desirable. One possibility would be to show energy-ordered profiles of ensemble properties as a function of ensemble size, since this reveals how representative the actual choice of ensemble is.[48]

Yet, none of these measures really relates to accuracy, i.e. the similarity of the calculated structure to the "true" structure. One can determine, however, whether a calculated structure is consistent with experimental data *not* used to constrain it. One such approach is cross-validation. A proportion of constraints is omitted from the structure calculation, and the consistency of the resulting structure with the unused constraints is taken as a measure of accuracy. (This is analogous to the procedures used by crystallographers in measuring $R_{(free)}$.) In NMR, however, constraints are sparse, so one cannot afford to leave out a sizeable proportion. Non-NOE data that have been suggested for assessing structural quality include chemical shifts, coupling constants and, most recently, residual dipolar couplings measured in weakly ordered media.[49] In each case, values can be predicted for a given model and compared to experimental values. There is, however, always a choice between holding back some data for use in validation and using all the data to calculate a more constrained structure. To date, the trend has been towards the latter.

Conclusions

Two factors dominate current developments in bioinformatics: (1) The amount of raw data is increasing in quantity, spectacularly so, and in quality. Methods for annotation are improving but by no means at a comparable rate. Tools for identification of errors are improving both through enhanced understanding of what to expect and from a better statistical base from which to flag outliers. (2) A proliferation of web sites provides different views or slices or means of access to these data; and an increasingly dense reticulation of these sites provides links among databanks and information-retrieval engines. These links provide useful avenues to applications,

unfortunately they also provide routes for propagation of errors in raw or immature data. Such errors in data or annotation are often subsequently corrected in the databanks but the corrections not passed on.

These observations have several implications:

Annotation is a weak component of the enterprise. Automation of annotation is possible only to a limited extent and getting annotation right remains labor-intensive. But the importance of proper annotation, however, cannot be underestimated. P. Bork has commented that for people interested in analysing the protein sequences implicit in genome sequence information, errors in gene assignment vitiate the high quality of the sequence data.

The only possible solution is a *distributed* and *dynamic* error-correction and annotation process. The workload must be distributed because databank staff have neither the time nor the expertise for the job; specialists will have to act as curators. The process must be dynamic, in that progress in automation of annotation and error identification/correction will permit reannotation of databanks. As a result, we will have to give up the "safe" idea of a stable databank composed of entries that are correct when they are first distributed in mature form and stay fixed thereafter. Databanks will become a seething broth of information both growing in size, and maturing—we must hope—in quality.

This will create problems, however, in organizing applications. Many institutions maintain local copies of databanks: At present, "maintain" means "top up;" yet this will no longer be sufficient. In the face of dynamically changing databanks, how can we avoid proliferation of various copies in various states? How will it be possible to reproduce a scientific investigation based on a database search? One possible solution is to maintain adequate history records in each databank itself in order to be able to reconstruct its form at any time. This is analogous to the information in the Oxford English Dictionary, which permits reconstruction of a English dictionary appropriate for 1616 or 1756.

To recover from the dispersion of outdated and/or erroneous information and links, perhaps "knowbots" will come to our rescue.[50] Knowbots are mobile software agents ("daemons") designed to cruise the internet. They could perform a continuous checking of the information resources for molecular biology.

It is also clear that the entire molecular biological community must become involved in the data archiving process. Experts will have to curate the collections and software engineers will have to organize the updating and access or distribution.

The community as a whole will have to play an active role in making intelligent decisions about what to archive. For example, at present the deposition of structure factors associated with atomic coordinates determined by X-ray crystallography is not mandatory. Deposition of structure factors would help resolve some of the problems in distinguishing real outliers from errors in structures. As the software for crystal-structure determination improves, it will be possible, at least in some cases, to redetermine the structure from the experimental data to produce more accurate results. In addition, new challenges are on the horizon. Planning for data archiving in the emergent proteome project—a dynamic analysis of spatiotemporal expression and activity of proteins in an organism—is one such challenge.

In the end, we will get the resources that we deserve. The question is whether we are willing to make the requisite effort to ensure the adequacy of their standards of quality.

Acknowledgments

References

1. See http://www.ebi.ac.uk/Databases/index.html.
2. Bashford D, Chothia, C, Lesk AM. Determinants of a protein fold: unique features of the globin amino acid sequences. J Mol Biol 1987;196: 199–216.
3. Bork P, Bairoch A. Go hunting in sequence databases but watch out for the traps. Trends Genet 1996;12:425–427.
4. Brändén C-I, Jones TA. Between objectivity and subjectivity. Nature 1990;343:687–689.
5. Korning PG, Hebsgaard SM, Rouzé P, Brunak S. Cleaning the GenBank *Arabidopsis thaliana* data set. Nucl Acids Res 1996;24:316–320.
6. Learn GH Jr, Korber BTM, Foley B, Hahn BH, Wolinsky SM, Mullins JI. Maintaining the integrity of human immunodeficiency virus sequence databases. J Virol 1996;70:5720–5730.
7. Lesk AM, Boswell DR, Lesk VI, Lesk VE, Bairoch A. A cross-reference table between the protein data bank of macromolecular structures and the national biomedical research foundation protein identification resource amino acid sequence data bank. Protein Sequences and Data Analysis 1989;2:295–308.
8. Kristensen T, Lopez R, Prydz SH. An estimate of the sequencing error frequency in the DNA sequence databases. DNA Sequence 1992;2:343–346.
9. Chen WQ, Hunkapiller T. Sequence accuracy of large DNA sequencing projects. DNA Sequence 1992;2:335–342.
10. Sulston J, Du Z, Thomas K, Wilson R, Hillier L, Halloran N, Green P, Thierry-Mieg J, Qiu L, Dear S, Coulson A, Craxton M, Durbin R, Berks M, Metzstein M, Ainscough R, Waterston R. The *C. elegans* genome sequencing project: a beginning. Nature 1992;356:37–41.
11. Khurshid F, Beck S. Error analysis in manual and automated DNA sequencing. Anal Biochem 1993;208:138–143.
12. Richterich P. Estimation of errors in "raw" DNA sequences: a validation study. Genome Res 1998;8:251–259.
13. Churchill GA, Waterman MS. The accuracy of DNA sequences: estimating sequence quality. Genomics 1992;14:89–98.
14. Lawrence CB, Solovyev V. Assignment of position-specific error probability to primary DNA sequence data. Nucl Acids Res 1994; 22:1272–1280.
15. Lipshutz RJ, Traverner F, Hennessy K, Hartzell G, Davis R. DNA sequence confidence estimation. Genomics 1994;19:417–424.
16. Bonfield JK, Staden R. The application of numerical estimates of base calling accuracy to DNA sequencing projects. Nucl Acids Res 1995;23:1406–1410.
17. Ewing B, Hillier L, Wendl MC, Green P. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. Genome Res 1998;8:175–185.
18. Ewing B, Green P. Base-calling of automated sequencer traces using phred. II. Error probabilities. Genome Res 1998;8:186–194.

19. Posfai J, Roberts RJ. Finding errors in DNA sequences. Proc Natl Acad Sci USA 1992;89:4698–4702.
20. Needleman SB, Wunsch CD. A general method applicable to the search for similarities in the amino acid sequence of two proteins. J Mol Biol 1970;48:443–453.
21. Pearson WR, Lipman DJ. Improved tools for biological sequence comparison. Proc Natl Acad Sci USA 1988;85:2444–2448.
22. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol 1990;215:403–410.
23. States DJ. Molecular sequence accuracy: analysing imperfect data. Trends Genet 1992;8:52–55.
24. http://www.ncbi.nlm.nih.gov/Web/Newsltr/feb98.html#GenBank.
25. White O, Dunning T, Sutton G, Adams M, Venter JC, Fields C. A quality control algorithm for DNA sequencing projects. Nucl Acids Res 1993;21:3829–3838.
26. Frishman D, Mironov A, Mewes H-W, Gelfand M. Combining diverse evidence for gene recognition in completely sequenced bacterial genomes. Nucl Acids Res 1998;26:2941–2947.
27. Fu DJ, Tang K, Braun A, Reuter D, Darnhofer-Demar B, Little DP, O'Donnell MJ, Cantor CR, Koster H. Sequencing exons 5 to 8 of the p53 gene by MALDI-TOF mass spectrometry. Nature Biotechnol 1998;16:381–384.
28. Marquart M, Deisenhofer J, Huber R, Palm W. Crystallographic refinement and atomic models of the intact immunoglobulin molecule Kol and its antigen-binding fragment at 3.0 (Å) and 1.9 (Å) resolution. J Mol Biol 1980;141:369–391.
29. Kleywegt GJ, Jones TA. Where freedom is given, liberties are taken. Structure 1995;3:535–540.
30. Kleywegt GJ, Brünger AT. Checking your imagination: applications of the free R value. Structure 1996;4:897–904.
31. Brünger AT. Free R-value: a novel statistical quantity for assessing the quality of crystal structures. Nature 1992;355:472–475.
32. Murshudov GN, Dodson EJ. Simplified error estimation à la Cruickshank in macromolecular crystallography. CCP4 Newsletter. 1997;
33. Tickle IJ, Laskowski RA, Moss DS. Error estimates of protein structure coordinates and deviations from standard geometry by full-matrix refinement of γB- and βB2-Crystallin. Acta Cryst 1998;D54:243–252.
34. Ohlendorf DH. Accuracy of refined protein structures. II. Comparison of four independently refined models of human interleukin-1β. Acta Cryst 1994;D50:808–812.
35. Luzzati V. Traitement statistique des erreurs dans la determination des structures crystallines. Acta Cryst 1952;5:802–810.
36. Cruickshank DWJ. Remarks about protein structure precision. Acta Cryst 1999;55:583–601.
37. Vaguine AA, Richelle J, Wodak SJ. SFCHECK: A unified set of procedures for evaluating the quality of macromolecular structure factor data, their agreement with the atomic model. Acta Cryst 1999;D55:191.
38. www.sdsc.edu/CCMS.
39. MacArthur MW, Laskowski RA, Thornton JM. Validation of protein models derived from experiment. Curr Opin Struct Biol 1994;4:731–737.
40. Kleywegt GJ, Jones TA. Model-building and refinement practice. Meth Enzymol 1997;277:208–230.
41. EU 3-D Validation Network. Who checks the checkers? Four validation tools applied to eight atomic resolution structures. J Mol Biol 1998;276:417–436.
42. Weissig H, Bourne PE. An analysis of the protein data bank in search of temporal and global trends. Bioinformatics 1999;15:807–831.
43. Bowie JU, Lüthy R, Eisenberg, D. A method to identify protein sequences that fold into a known three-dimensional structure. Science 1991;253:164–170.
44. Kleywegt GJ, Jones TA. Phi/psi-chology: Ramachandran revisited. Structure 1996;4:1395–1400.
45. Wüthrich K. NMR of proteins and nucleic acids, John Wiley, New York. 1986;
46. Peng JW, Wagner G. Investigations of protein motions via relaxation measurements. Meth Enzymology 1994;239:563–596.
47. Doreleijers JF, Rullmann JA, Kaptein R. Quality assessment of NMR structures: a statistical survey. J Mol Biol 1998;281:149–164.
48. Fletcher CM, Jones DNM, Diamond R, Neuhaus D. Treatment of NOE constraints involving equivalent or nonstereoassigned protons in calculations of biomacromolecular structures. J Biomolec NMR 1996;8:292–310.
49. Tjandra N, Bax A. Direct measurement of distances and angles in biomolecules by NMR in a dilute liquid crystal medium. Science 1997;278:1111–1114.
50. Lesk M Practical Digital Libraries: Books, Bytes and Bucks. San Francisco: Morgan Kaufmann, 1997; p. 25.
51. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997;25:3389–3402.
52. Kleywegt GJ, Jones TA. Databases in protein crystallography. Acta Cryst 1998;D54:1119–1131.
53. Ramachandran GN, Ramakrishnan C, Sasisekharan V. Stereochemistry of polypeptide chain configurations. J Mol Biol 1963;7:95–99.
54. Kleywegt GJ, Hoier H, Jones TA. A re-evaluation of the crystal structure of chloromuconate cycloisomerase. Acta Cryst 1996;D52:858–863.
55. Hoier H, Schlomann M, Hammer A, Glusker JP, Carrell HL, Goldman A, Stezowski JJ, Heinemann U. Crystal structure of chloromuconate cycloisomerase from Alcaligenes eutrophus JMP134 (pJP4) at 3 (Å) resolution. Acta Cryst 1994;D50:75–84.