



Biotechnology

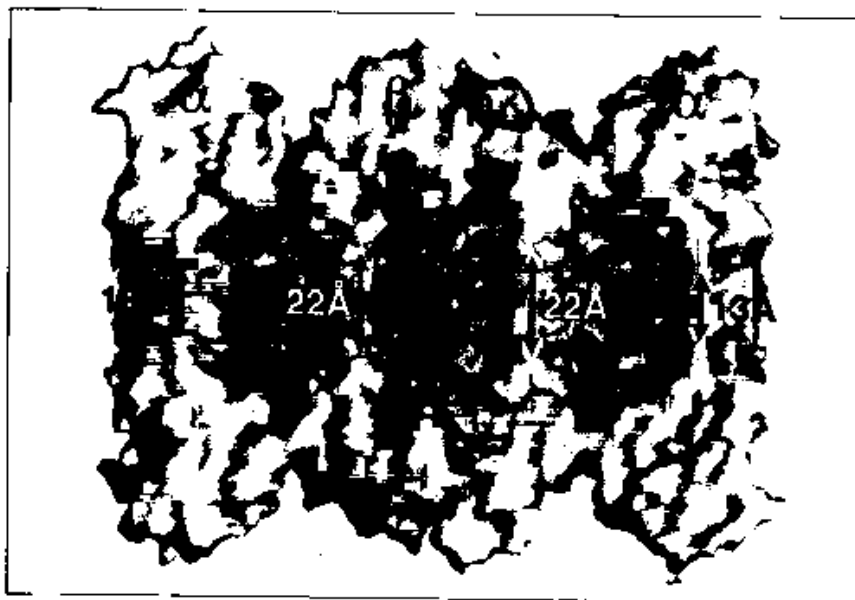
Protein engineering

Edited by Abraham M de Vos and Andreas Plöckhün

Commentary

Pitfalls of protein sequence analysis

World Wide Web sites



Pitfalls of protein sequence analysis

Commentary

Burkhard Rost* and Alfonso Valencia†

Addresses

*European Molecular Biology Laboratory, Protein Design Group, 69012 Heidelberg, Germany; e-mail: rost@embl-heidelberg.de

†CNB-CSIC Cantoblanco, 28049 Madrid, Spain; e-mail: valencia@samba.cnb.uam.es

Current Opinion in Biotechnology 1998, 7:457-461

© Current Biology Ltd ISSN 0959-1689

Abbreviations

1D one-dimensional
3D three-dimensional
EST expressed sequence tag

Introduction

Imagine you have a sequenced protein, either sequenced in your own lab or pulled down from genome projects of expressed sequence tag (EST) production. You decide to let theoretical biology assist you in finding *a priori* information about your protein that may be useful to accelerate and design experiments. You submit your sequence to database-search and/or structure-prediction services. The possible pitfalls are numerous, including picking a lousy server or misinterpreting the results. We give examples of common pitfalls experienced after 80 000 requests to an automatic prediction service (Table 1; [1]).

What can theory predict of protein structure?

In general, protein three-dimensional (3D) structure cannot be predicted from sequence [2,3]. However, 3D structure can be predicted by homology modelling, that is, by using a sequence homologue (>25% sequence identity) with an experimentally determined 3D structure. If no sequence homologue is found in the Protein Data Bank [4], there still is a chance of predicting 3D structure by threading, that is, by remote homology modelling (<25% sequence identity). However, correct 3D models — and even correct detection of remote homology — from threading are rare [5,6]. But theory can assist by predicting one-dimensional (1D; [2]) aspects of 3D structure, for example, secondary structure, solvent accessibility, trans-membrane helices, binding sites, sequence motifs, and aspects of protein function.

Ease of use bears an ease of misuse

Rapidly developing electronic communication (Internet, World Wide Web) facilitates the spreading of prediction methods. Experimental biologists submit sequences and theoretical biologists configure automatic services that return predictions. The advantage is that users need not become experts in the use of sequence-analysis tools.

However, the ease of offering and accessing predictions carries two problems: inaccurate methods (or insufficiently validated ones) are made available that bypass selection systems such as referees; and users may misinterpret results because of a lack of insight into the features of prediction methods.

Sequence alignments

More than 30% pairwise sequence identity

Sequence analysis usually begins with a search for homologues in databases [4,7]. The success of alignment programs relies on evolutionary connections between homologous proteins: if 24 out of 80 aligned residues (i.e. 30%; more for shorter matches [8]) are identical between two naturally evolved proteins, the two have similar 3D structures and similar functions [8,9,10] (this may not be valid for engineered proteins). The level of sequence identity significant for homology is much higher for smaller regions; for very short motifs (e.g. RGD, KDEE, [single letter amino acid code]) homology cannot be inferred from sequence identity.

Higher values for sequence similarity

If similarity scores (physicochemical properties; D→E = 1) rather than identity scores (D→E = 0, D→D = 1) are used to select homologues, the pairwise similarity usually has to be higher than 30% to be significant. A rule of thumb for true homologues is that for these, similarity scores are higher than identity scores. Similarity scores depend on the particular similarity metric used. Thus, results cannot be compared directly between different methods.

Constraints to significant identity: composition bias and gaps

There are two possible errors in inferring homology from a given level of pairwise sequence identity. The first involves composition bias: if the two aligned proteins have regions with a high composition of certain amino acids (e.g. ARG-rich regions in DNA-binding proteins), such regions may be important for protein function — and in many cases are indicative of functional class — but may be misleading for homology searches. Thus, composition biased regions should be ignored when compiling sequence identity, and be used only to confirm presence of similar composition bias in identified homologues. The second involves the presence of many gaps: if an alignment between two proteins contains too many insertions (gaps), even a relatively high value of sequence identity may not suffice to ascertain homology (typical structure alignments contain up to 10% gaps).

Table 1

Common pitfalls and ways to avoid them.

Pitfall	Cause	Result	Fix
Overinterpretation of sequence similarity	Level of similarity too low (<30%) Too many gaps Similarity confused with sequence identity Composition bias	Incorrectly inferred homology and hence wrong function or wrong structure	Use thresholds considering gaps, similarity and composition bias
Insufficient description of family	Full family not matching local motifs	Pattern/homology incorrectly inferred	Use profiles of subfamily Repeat search with different family members
Inaccurate function designation in database	Errors in original paper Wrong annotation in database Overinterpretation of homology in database	Incorrectly inferred function Too detailed predictions for function	Check literature Compare several homologues Compare with level of divergence in family
Overinterpretation of secondary structure prediction	Single residue assignments taken too literally	Wasted time Incorrectly inferred function	Consider mainly segment level
Overinterpretation of threading	Trendy False prediction Prediction not reliable	Wrong or partially wrong 3D model	Use controls appropriate to particular threading method Check functional residues
Overinterpretation of 3D model	3D model built on shaky alignment information, or unreliable structure Large loops modelled Wrong choice of modelling and/or checking tools	Wrong interpretation of location of residue in 3D structure Wrong prediction of structure-function relationship	Do more careful alignment Take into account reliability index for homology models Low level similarity → coarse-grained tools High level similarity → full-atom description tools

Evolutionary patterns crucial for successful prediction of function

A typical mistake is to predict function by putative homology on the basis of an overinterpreted level of sequence similarity. Functional and structural constraints are translated into sequence conservation in a particular way that depends on the particular protein structure and its evolution. The level of similarity required for identifying functionally equivalent proteins in two species depends on the overall divergence of the species and on the particular protein family.

Some databases use more reliable annotations than others

When predicting function on the basis of a protein's similarity to proteins of known function (as annotated in databases), it is important to be aware of incomplete or wrong annotations. The annotations for the putative homologue ought to be verified in the original sources of the functional assignments (a more reliable database is SWISS-PROT [7]). A similar problem arises for errors in sequences, such as frame shifts or sequencing errors (very frequent in ESTs).

Quality of alignment

Despite the central role that alignment programs play in sequence analysis, a thorough analysis of the quality of methods using statistically significant numbers of proteins

has yet to be accomplished. In general, alignments are more likely to be correct for higher levels of pairwise sequence identity, and are less likely to be correct in more variable regions.

Stability of alignment

Say you find three proteins you want to use to build a multiple alignment for your sequence of unknown structure (U). In experiment A, you align them in the order 1-U (the first protein with U), 2-U, 3-U; in experiment B you inverse the order to get 3-U, 2-U, 1-U. The alignments of A and B may differ in detail. Is this an error of the program? Not necessarily; the reason may be that the alignment is just not unique, that is, the best and the second best solution to the alignment problem may have similar scores. In such cases, the alignment is less reliable in regions where the results from A and B differ.

Sequence alignments reveal underlying evolutionary processes

Aligning protein sequences may appear to be purely a problem of matching letters. However, sequence alignments unravel information about structural and functional relations between residues in different proteins. Obviously, it is not trivial to map the complexity of factors determining protein structure and function onto 1D relations between letters.

Evolutionary divergence within sequence families

In general, regions with many insertions and deletions in the alignment are less informative. To illustrate this, say your protein has 333 residues; 22 sequences are aligned in the amino-terminal region, and only two near the carboxyl terminus. Then you cannot draw firm conclusions about function and/or structure from the conservation patterns at the carboxyl terminus. Do 20 sequences suffice for an informative alignment? Not necessarily. The information contained in a multiple alignment is determined by the divergence of the aligned sequences rather than by the number. Ideally, the entire range between 30 and 90% sequence identity should be covered, preferably with many sequences at lower levels (30–50%).

Extending local sequence motifs to entire folds

The goal of database searches is to find a good alignment for a full folding domain. This task is often very difficult in the absence of 3D information. If you found a local motif (e.g. by a BLAST search [11]), you would try to extend the alignment to cover the full core of the proteins. A good indication of a correct match is that motifs (or local hits from the BLAST search) appear in the same order in the final alignment.

Aligning entire families rather than subsets of sequences

Another helpful criterion indicative of true homologues is that the 'full' protein family is described by the alignment rather than just a subset of sequences. In practice, searches often initially identify only a few members of a given family. The aligned regions should then be investigated thoroughly: are local motifs compatible with the entire family? Are there other motifs that could be used to uncover further homologues by restricting the search to such motifs? Is the pattern symmetrical, that is, if your protein U has been aligned to, for example, the protein kinase family on the basis of strong local motifs, do other patterns relevant for the kinase family match in U? Incomplete family alignments are often indicative of a misleading local pattern and of having falsely aligned unrelated proteins.

Profile alignments may intrude into the twilight zone

In the twilight zone [10] of 20–30% pairwise sequence identity, sequence alignments become tricky. Only methods using profiles derived from the sequence family of your protein U may reliably intrude into that zone. The quality of such alignments depends crucially on the information contained in the alignment, that is, the size (number of sequences) and divergence (levels of pairwise sequence identity) of the sequence family. In sparse regions (less sequences), alignments are generally less reliable. However, penetrating the twilight zone requires attention!

Predictions of 1D protein structure**70% correct implies 30% incorrect**

The most accurate methods for predicting secondary structure or solvent accessibility are based on multiple alignment information and reach levels of about 70% accuracy. This level of accuracy suffices to render useful predictions [6,12]. However, in interpreting the predictions, it is often instructive to spot the 30% of residues you suspect to be falsely predicted.

Spread of prediction accuracy

An expected accuracy of 70% does not imply that for your protein U, 70% of all residues are correctly predicted. Instead, values published for prediction accuracy are averaged over hundreds of unique proteins. An expected accuracy of $70 \pm 10\%$ (one standard deviation) implies that, on average, for two thirds of all proteins, between 60 and 80% of the residues will be predicted correctly. Thus, prediction accuracy can be higher than 80% or lower than 60% for your protein.

Special classes of proteins

Prediction methods are usually derived from knowledge contained in subsets of proteins from databases. Consequently, they should not be applied to classes of proteins that have not been included in the subsets. For example, methods for predicting helices in globular proteins are likely to fail when applied to predicting transmembrane helices. In general, results should be taken with caution for proteins with unusual features, such as proline-rich regions, an unusually high number of cysteine bonds, or for domain interfaces.

Better alignments yield better predictions

Multiple alignment based predictions are substantially more accurate than single sequence based predictions. How many sequences do you need in your alignment to expect an improvement, and how sensitive are prediction methods with respect to errors in the alignment? The more divergent sequences contained in the alignment the better (two distantly related sequences often improve secondary structure predictions by several percentage points). Regions with few aligned sequences yield less reliable predictions. The sensitivity to alignment errors depends on the methods, for example, secondary structure prediction is less sensitive to alignment errors than accessibility prediction.

Better + worse = even better?

Today, several automatic services accomplish secondary structure predictions. Some users fall into the 'what is common is correct' trap, that is, they average over all prediction methods and consider identical regions as more reliable. In exceptional cases, such a majority vote may be beneficial; however, frequently, the result will be the worst-of-all prediction. Often, it is preferable to use reliability indices provided by some methods. Such indices

answer the question: how reliably is the tryptophan at position 307 predicted in a surface loop? (Note that the correlation between such indices and prediction accuracy is sufficiently tested for a few methods only.)

1D structure may or may not be sufficient to infer 3D structure

Say you obtain the following prediction for regular secondary structure: helix-strand-strand-helix-strand-strand (H-E-E-H-E-E), then assume that you find a protein of known structure with the same motif. Can you conclude that the two proteins have the same fold? The answer is yes and no: your guess may be correct, but there are various ways to realise the given motif by completely different structures. For example, the secondary structure motif H-E-E-H-E-E is contained in at least 16 structurally unrelated proteins.

Predictions of 3D structure

Accuracy of homology modelling at the level of ribbon plots

A common mistake in using homology-derived models of 3D structure is that the model is taken too literally. In general, the accuracy of homology modelling decreases with lower levels of pairwise sequence identity between your sequence U and the target structure. Models accurate enough to simulate ligand binding in detail require levels of above 90% pairwise sequence identity. Furthermore, successful applications of homology modelling may be hampered by two other difficulties: firstly, loop regions are, in general, less reliable; and secondly, accurate predictions for regions with insertions or deletions, that is, regions where the template structure does not match U, are the exception [6].

Avoid overinterpreting the details of the model

The main purpose of homology modelling is to translate a given alignment into more intuitive 3D images. Such images often look temptingly 'real'. It is crucial to bear in mind the regions that were unreliable in the alignment or in the original structure (e.g. flexible regions, high R factor, low number of NMR constraints). Homology modelling tends to yield sketches of structure rather than accurate coordinates. Is homology modelling of any use? Is any hypothesis about a structure better than no hypothesis? Guessing details about protein function is such a difficult task that any null hypothesis may help to guide experiments. This may result in overinterpretations of the level of homology suggested by the sequence alignment and the level of accuracy of the resulting model. In practice, modelling often strikes a balance between users pushing for more interpretations and models pinpointing the limitations of the methods.

Remote homology modelling (threading) is extremely tricky!

One problem of homology modelling for lower levels of pairwise sequence identity is to get the alignment

between your sequence U and the template structure T correct. But even if the alignment were correct, a principle limitation is that T and U just do not have identical 3D structures. This problem is particularly fatal for remote homology modelling (threading), that is, the prediction of 3D structure based on less than 25% pairwise sequence identity. However, current threading methods are even more limited: getting the alignment correct is the exception rather than the rule [6]. The basic message of this statement is not 'don't use threading programs', but 'use them with extreme caution and be aware that most resulting models are likely to be mostly wrong'.

Check the predicted model

No matter which homology modelling technique you use, you'd better check the model by one of the various programs that detect errors in experimentally determined structures [2]. What if all checks reveal your model to resemble a known structure? Then you have a good starting point. But keep in mind that checking tools are tailored to spot errors in structures derived from NMR or X-ray crystallography. Your model may have been subject to extensive refinement trying to optimize exactly those variables that are checked (e.g. torsion angles, bond lengths). Furthermore, other features that help to spot errors in experimentally determined structures (e.g. insertion of gaps in secondary structure elements, proximity of functional or active site residues, hydrophobicity of core) may have been already avoided for your model by the alignment program. What if the checks reveal that your model is not native-like? Shifts of secondary structure segments are rather frequent in protein evolution. It may have been misleading to optimize bumps or side-chain packing by the modelling software in the first place when the deviations between the secondary structure elements of the target and your protein were substantial. Try to focus on more reliable regions and/or particular aspects of the model compatible with independently derived information.

Conclusions

Most mistakes are pitfalls that could have been avoided

Most of the examples listed can be traced in the literature. Is the take-home message 'hands off applying prediction methods if you are not an expert'? Certainly not: the pitfalls listed can be avoided. The more difficult the prediction task, the more skills are required: threading and homology modelling are difficult, alignments and 1D structure prediction straightforward. Understanding the limitations of the tools is the key to a successful application.

Is your protein a suitable target for prediction methods?

In answer to this question, the general message from a workshop organized by Anna Tramantano (Istituto di Ricerche di Biologia Molecolare, Rome) and Tim Hubbard (Medical Research Council, Cambridge) [12] was the following: this depends on how much you are interested

in finding answers to your questions! Theoretical biology still fails to predict 3D structure from sequence, but predictions of various simplified 1D aspects of structure become more accurate and more useful with every new sequence added to public databases.

References

1. Rost B: PHD: predicting one-dimensional protein structure by profile based neural networks. *Methods Enzymol* 1996, 246:525-539.
2. Rost B, Sander C: Structure prediction of proteins - where are we now? *Curr Opin Biotechnol* 1994, 6:372-380.
3. Rost B, Sander C: Bridging the protein sequence-structure gap by structure predictions. *Annu Rev Biophys Biomol Struct* 1996, 25:in press.
4. Bernstein FC, Koetzle TF, Williams GJB, Meyer EF, Brice MD, Rodgers JR, Kennard O, Shimanouchi T, Tasumi M: The Protein Data Bank: a computer based archival file for macromolecular structures. *J Mol Biol* 1977, 112:535-542.
5. Shortle D: Protein fold recognition. *Nat Struct Biol* 1995, 2:91-92.
6. Littman EE: Protein structure prediction: a special issue. *Proteins* 1995, 23:295-460.
7. Bairoch A, Apweiler R: The SWISS-PROT protein sequence data bank and its new supplement TREMBL. *Nucleic Acids Res* 1995, 24:21-25.
8. Sander C, Schneider R: Database of homology-derived structures and the structural meaning of sequence alignment. *Proteins* 1991, 9:56-58.
9. Chothia C, Lesk AM: The relation between the divergence of sequence and structure in proteins. *EMBO J* 1986, 5:823-828.
10. Feng D-F, Doolittle RF: Progressive sequence alignment as a prerequisite to correct phylogenetic trees. *J Mol Evol* 1987, 25:351-360.
11. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: Basic local alignment search tool. *J Mol Biol* 1990, 215:403-410.
12. Hubbard T, Tramontano A, The 1995 IRBM Workshop Team: Update on protein structure prediction: results of the 1995 IRBM workshop. *Folding Des* 1996, 1:R55-R63.