Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

ELSEVIER

Journal of Biotechnology 36 (1994) 185–220

journal of biotechnology

Review

The blind watchmaker and rational protein engineering

Henrik W. Anthonsen, Antônio Baptista, Finn Drabløs, Paulo Martel, Steffen B. Petersen *

MR-Center, SINTEF UNIMED, N-7034 Trondheim, Norway

Received 7 March 1994; accepted 23 April 1994

Abstract

In the present review some scientific areas of key importance for protein engineering are discussed, such as problems involved in deducting protein sequence from DNA sequence (due to posttranscriptional editing, splicing and posttranslational modifications), modelling of protein structures by homology, NMR of large proteins (including probing the molecular surface with relaxation agents), simulation of protein structures by molecular dynamics and simulation of electrostatic effects in proteins (including pH-dependent effects). It is argued that all of these areas could be of key importance in most protein engineering projects, because they give access to increased and often unique information. In the last part of the review some potential areas for future applications of protein engineering approaches are discussed, such as non-conventional media, de novo design and nanotechnology.

Key words: Protein engineering; Protein sequence; Homology; NMR; Molecular dynamics; Protein electrostatics

1. Introduction

Nature has evolved using several types of random mutations in the genetic material as a fundamental mechanism, thereby creating new versions of existing proteins. By natural selection Nature has given a preference to organisms with proteins which directly or indirectly made them better adapted to their environment. Thus Nature works like a blind watchmaker, trying out an endless number of combinations. This may seem to be an inefficient approach by industrial standards, but nevertheless Nature has been able to develop some highly complex and sophisticated designs, simply by the power of natural selection over millions of years, occurring in a large number of parallel processes. By virtue of reproduction several copies of each organism have been able to test the effect of different mutations in parallel. It is quite probable that the mutation frequency was higher in ancient species (Doolittle, 1992), although it is still possible to find highly mutable loci in genes involved in adaptation to the environment (Moxon et al., 1994).

Enzymes have been used by man for thousands of years for modification of biological molecules. The use of rennin (chymosin) in rennet for cheese production is a relevant example. And with increased knowledge about proteins, genes and other biological macromolecules scientists started

* Corresponding author.

Elsevier Science B.V.
SSDI 0168-1656(94)00063-I

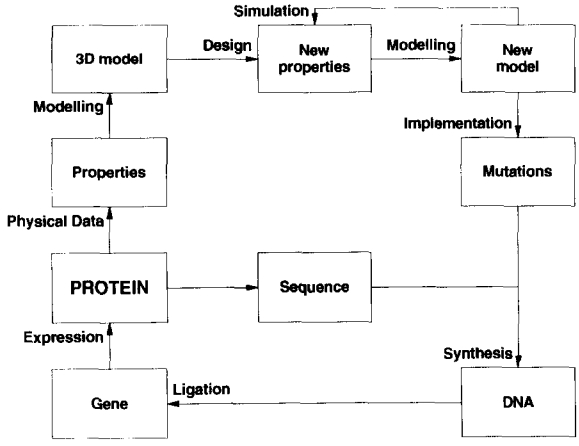


Fig. 1. The protein engineering process. Starting with a protein with known sequence and properties, we make a 3-D model of the protein from experimental structure data or by homology. By modelling and simulation we identify mutations that will modify selected properties of the protein (the *design* part of the process), these mutations are implemented at the DNA level and expressed in a suitable organism (the *production* part of the process), and the success of the design is verified by experimental methods.

to look at methods for making modified proteins with new or improved properties. At first this was done by speeding up Nature's own approach, by increasing the number of mutations (e.g., by using chemicals or radiation) and by using a very strong selection based on tests for specific properties.

With the introduction of new and powerful techniques for structure determination and site directed mutagenesis, it is now possible to do rational protein modification. Rather than testing out a large number of random mutations, it has become feasible to identify key residues within the protein structure, to predict the effect of changing these residues, to implement these changes in the genetic material, and finally to produce large amounts of modified proteins. This is protein engineering.

There are several reviews describing the fundamental ideas in protein engineering, see Fersht and Winter (1992) for a recent one. The basic protein engineering process is shown in Fig. 1 (see also Petersen and Martel (1994)). In most cases it starts out with an unmodified protein with well-characterised properties. For some reason we want to modify this protein. In the case of an enzyme we may want to make it more stable, alter the specificity or increase the catalytic activity. First we enter the design part of the protein engineering process. Based on structural data we create a computer model of the protein. By a combination of molecular modelling and experimental methods the correlation between relevant properties and structural features is established, and changes affecting these properties can be identified and evaluated for implementation. In more and more cases the effect of these changes can be simulated, and the modifications can be optimised with respect to these simulations.

As soon as a new design has been established we may enter the production part of the process. The necessary mutations must be implemented in the genetic material, this genetic material is introduced into a production organism, and the resulting modified protein can (in most cases) be extracted from a bioprocess. This protein can be tested with respect to relevant properties, and if necessary it may be used as a basis for re-entering the design part of the protein engineering process. After a few iterations we may reach an optimal design.

There are several examples of successful protein engineering projects. Protein engineering may be used to improve protein stability (Kaarsholm et al., 1993), enhance or modify specificity (Getzoff et al., 1992; Witkowski et al., 1994), adapt proteins to new environments (Arnold, 1993; Gupta, 1992), or to engineer novel regulation into enzymes (Higaki et al., 1992). In some cases even de novo design of new proteins may be relevant, using knowledge gained from existing structures (Kamtekar et al., 1993; Johnson et al., 1993; Shakhnovich and Gutin, 1993; Ghadiri et al., 1993; Ball, 1994).

In a truly multidisciplinary project chymosin mutants with optimal activity at increased pH values compared to wild-type chymosin was designed and produced (Pitts et al., 1992). Point mutations changing the charge distribution of superoxide dismutase have been used to increase reaction rate by improved electrostatic guidance (Getzoff et al., 1992). A project on converting trypsin into chymotrypsin has been important for understanding the role of chymotrypsin surface

loops (Hedstrom et al., 1992), a serine active site hydrolase has been converted into a transferase by point mutations (Witkowski et al., 1994), and mutations in insulin aiming at increased folding stability have given an insulin with enhanced biological activity (Kaarsholm et al., 1993).

An example of a rational de novo project (as opposed to the random approach used, e.g., in generation of catalytic antibodies) is the design of an enzymatic peptide catalysing the decarboxylation of oxaloacetate via an imine intermediate, in which a very simple design gave a three to four orders of magnitude faster formation of imine compared to simple amine catalysts (Johnson et al., 1993).

In some cases it may also be an interesting approach to incorporate nonpeptidic residues into otherwise normal proteins (Baca et al., 1993), or to build de novo proteins by assembling peptidic building blocks on to a nonpeptidic template (Tuchscherer et al., 1992). It has been shown that incorporation of nonpeptidic residues into β-turns of HIV-1 protease gives a more stable enzyme (Baca et al., 1993). The main problem with this approach is how to incorporate the non-standard residues. In the HIV-1 protease case solid-phase peptide synthesis combined with traditional organic synthesis was used, others have suggested that the degeneracy of the genetic code may be used to incorporate novel residues via the standard protein synthesis machinery of the cell (Fahy, 1993).

In the present review we will look at the design part of the protein engineering process, with emphasis on some of the more difficult steps, especially homology based modelling in cases with very low sequence similarity, nuclear magnetic resonance (NMR) of very large proteins and modelling of electrostatic interactions. In the last part of the review we will discuss some possible future directions for protein engineering and protein design.

2. From DNA sequence to protein sequence – a non-trivial step

Any protein engineering project is based on information about the protein sequence. This information may stem from either direct protein sequencing or a deducted translation of the DNA/RNA sequence. The amount of information on protein and nucleic acid sequences, as well as on relevant data like 3-D structures and disease-related mutations, is growing at a very rapid pace, and novel databases and computer tools give increased access to these data (Coulson, 1993). It is very reasonable to expect that projects like the human genome project will succeed in providing us with sequence information about every single gene in our chromosomes within the next decade. This information will be

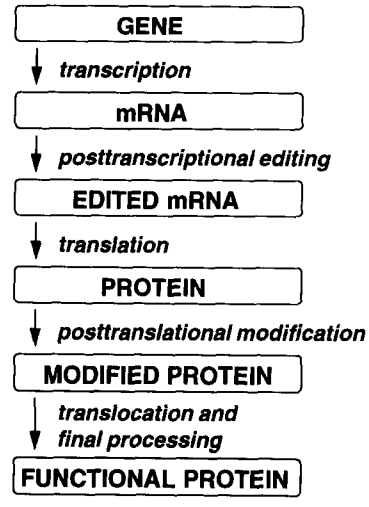


Fig. 2. Protein synthesis. The figure illustrates the different steps that may be involved in the synthesis of a functional protein. After transcription, the mRNA may be edited, a process that now has been reported in man, plants and primitive organisms (*Trypanosoma brucei*). The mRNA is then translated into a protein sequence. This protein sequence can subsequently be modified, leading to N- or O-glycosylation, phosphorylation, sulfation or the covalent attachment of fatty acid moieties to the protein. At this stage the protein is ready for transport to its final destination – which may be right where it is at the time of synthesis, but the destination may also be extracellular or in secluded compartments such as the mitochondria or lysosomes. In this case the protein is equipped with a signal sequence. After arrival to its destination the protein is processed, often involving proteolytic cleavage of the signal sequence. Shorter routes to the functional protein with fewer steps undoubtedly exist as well as routes with interchanged steps of processing. Finally, the catabolism of the protein is also part of the process, but has been left out in the figure.

of key importance for our understanding of the biology, development and evolution of man. It should, however, be kept in mind that the sequence itself may give us little information about regulation of gene expression, i.e., under what conditions genes are expressed, if they are expressed at all.

2.1. Posttranscriptional editing

Most protein sequences have been deducted from gene sequences. It is in most cases a priori assumed that a trivial mapping exists between the two sets of information. However, this may not necessarily be the case. In Fig. 2, the various steps currently recognised as being of importance for the production of the mature enzyme are shown, and several of these steps may affect the mapping from gene to protein. Posttranscriptional editing is modifications at the mRNA level affecting the mapping of information from gene to protein, often involving modification, insertion or deletion of individual nucleotides at specific positions (Cattaneo, 1994). Currently only speculative models exist for the underlying molecular mechanism(s) for posttranscriptional editing.

In the case of mammalian apolipoprotein B two forms exist, both originating from a single gene. The shorter form, apo B48, arises by a posttranscriptional mRNA editing whereby cytidine deamination produces an UAA termination codon (Teng et al., 1993).

In the AMPA receptor subunit GluR-B mRNA editing is responsible for changing a glutamine codon (CAG) into a arginine codon (CGG) (Higuchi et al., 1993). This editing has a pronounced effect on the Ca^{2+} permeability of the AMPA receptor channel, and it seems to be controlled by the intron–exon structure of the RNA. Similar mRNA editing has been reported in the related kainate receptor subunits GluR-5 and GluR-6, where two additional codons in the first trans-membrane region are altered (Sommer et al., 1991; Köhler et al., 1993). It is also interesting in this context that certain human genetic diseases have been related to reiteration of the codon CAG (Green, 1993).

mRNA editing in plant mitochondria and chloroplasts has also been reported (Gray and Covello, 1993). Here the posttranscriptional mRNA editing consists almost exclusively of C to U substitutions. Editing occurs predominantly inside coding regions, mostly at isolated C residues, and usually at the first or second position of the codons, thus almost always changing the amino acid compared to that specified by the unedited codon.

In *Trypanosoma brucei* some extensive and well-documented posttranscriptional cases of editing have been reported (Read et al., 1992; Harris et al., 1992; Adler et al., 1991). The editing takes place at the mitochondrial transcript level where a large number of uridine nucleic acid bases are added or deleted from the mRNA, which then subsequently is translated.

Several non-editing processes affecting the transcription/translation steps are also known. Although the ribosomes in an almost perfect manner translate the message provided by the mRNA (with error rates less than 5×10^{-4} per amino acid incorporated), it appears as if the mRNA in certain cases contain information, that forces the ribosome to read the nucleic acid information in a non-canonical fashion (Farabaugh, 1993). A special case, that may deserve some attention as well, is the seleno proteins, were seleno cystein is introduced into the protein by an alternative interpretation of selected codons (Böck et al., 1991; Yoshimura et al., 1990; Farabaugh, 1993). Translational frameshifting has been found in retroviruses, coronaviruses, transposons and a prokaryotic gene, leading to different translations of the same gene. Two cases of translational 'hops' have been reported, where a segment of the mRNA is being skipped by all ribosomes, in the two cases 50 and 500 nucleotides were skipped, respectively (Farabaugh, 1993).

To our knowledge posttranscriptional editing and related processes are uncommon but definitely present in humans. It is, therefore, important to understand precisely how these mechanisms work, in order to correctly deduct the protein sequence from the gene sequence.

2.2. Posttranslational modifications

The most common posttranslational modifications are side chain modifications like phosphorylations, glycosylations and farnesylations, as well as others. However, some modifications may also affect the (apparent) gene to protein mapping.

Posttranslational processing may involve removal of both terminal and internal protein sequence fragments. In the latter case an internal protein region is removed from a protein precursor, and the external domains are joined to form a mature protein (Hodges et al., 1992; Xu et al., 1993). Interestingly, all intervening protein sequences reported so far have sequence similarity to homing endonucleases (Doolittle, 1993), which also can be found in coding regions of group I introns (Grivell, 1994).

Posttranslational modifications like phosphorylation, glycosylation, sulfation, methylation, farnesylation, prenylation, myristylation and hydroxylation should also be considered in this context. They modify properties of individual residues and of the protein, and may thus make surface prediction, dynamics simulations and structural modelling in general more complex. The residues that are specifically prone to such modifications are tyrosines (phosphorylation and sulfation), serine and threonine (O-glycosylation), asparagine (N-glycosylation), proline and lysine (hydroxylation) and lysine (methylation). In addition glutamic acid residues can become γ-carboxylated leading to high affinity towards calcium ions (Alberts, 1983). Specific transferases are involved in the modification, e.g., tyrosylprotein sulfotransferases (Suiko et al., 1992) and farnesyl-protein transferases (Omer et al., 1993).

Phosphorylation of amino acid residues is an important way of controlling the enzymatic function of key enzymes in the metabolic and signalling pathways. Tyrosine kinases phosphorylate tyrosine residues – thus introducing an electrostatic charge at a residue, which under normal physiological pH is uncharged. Phosphorylation is central to the function of many receptors, such as the insulin and insulin-like growth factor I receptors.

Given the possibility that several modifications may be introduced in the sequence when we move from gene to mature protein, the task of deducting a protein sequence from the gene sequence may be more complex than we normally assume.

3. Experimental 3-D structure determination

Although the protein sequence itself is a valuable starting point, the optimal basis for a rational protein engineering project will be a full structure determination of the protein. In many cases this turns out to be an expensive and time-consuming part of the project.

Most structure determinations are based on X-ray crystallography. This approach may give structures of atomic resolution, but is limited by the fact that stable high quality crystals are needed. Many proteins are very difficult to crystallise, in particular many structural and membrane-associated proteins.

A large number of important X-ray structures have been published over the last few years, and the structures of the *Hha*I methylase (Klimasauskas et al., 1994), the TBP/TATA-box complex (Kim et al., 1993a; Kim et al., 1993b) and the porcine ribonuclease inhibitor (Kobe and Deisenhofer, 1993) are mentioned as examples only.

NMR may be an alternative in many cases, as the proteins can be studied in solution, and for some experiments they can even be membrane associated. However, NMR is limited to relatively small molecules, and even with incorporation of labelling in the protein the upper limit for a full structure determination using current state of the art methods seems to be close to 30 kDa. Some novel techniques for studying structural aspects of larger proteins will be discussed (vide infra).

Representative examples of important NMR structures may be interleukin 1β (Clore et al., 1991a), the glucose permease IIA domain (Fairbrother et al., 1992) and the human retionic acid receptor-β DNA-binding domain (Knegtel et al., 1993).

Cryo electron microscopy (CEM) is a relatively new approach to protein structure determination. The resolution of the structures are still lower

than the corresponding X-ray structures, and a 2-dimensional crystal is a prerequisite. However, despite this CEM appears to be a very promising approach to structure determination of membrane associated proteins that can form 2-dimensional crystals. CEM has been used to study the nicotinic acetylcholine receptor at 9 Å resolution (Unwin, 1993) and the ATP-driven calcium pump at 14 Å resolution (Toyoshima et al., 1993), and in a combined approach using high resolution X-ray data superimposed on CEM data the structure of the actin-myosin complex (Rayment et al., 1993) and of the adenovirus capsid (Stewart et al., 1993) has been studied. The recent structure by Kühlbrandt et al. (1994) of the chlorophyll *a*/*b*-protein complex at 3.4 Å resolution shows that the resolution of CEM rapidly is approaching the resolution of most X-ray protein data.

Scanning tunnelling microscopy (STM) is another new approach for studying protein structures (Amrein and Gross, 1992; Lewerenz et al., 1992; Haggerty and Lenhoff, 1993). The method is interesting because of a very high sensitivity, as individual molecules may be examined. The method will give a representation of the surface of the molecule, rather than a full structure determination. However, it is possible that both CEM and STM can be used for identification of protein similarity. If data from these methods show that the overall shape of a protein is similar to some other known high resolution protein structure, then the known structure may be evaluated as a potential template for homology based modelling. We believe that such a model can either be used as an improved starting point for a full structure determination (i.e., for doing molecular replacement on X-ray data), or as a low resolution structure determination by itself.

4. Homology based modelling

In homology based modelling a known structure is used as a template for modelling the structure of an homologous sequence, based on the assumption that the structures are similar. This is a very simple and rapid process, compared to a full structure determination. The sequences may be homologous in the strict sense, meaning that there is an evolutionary relationship between

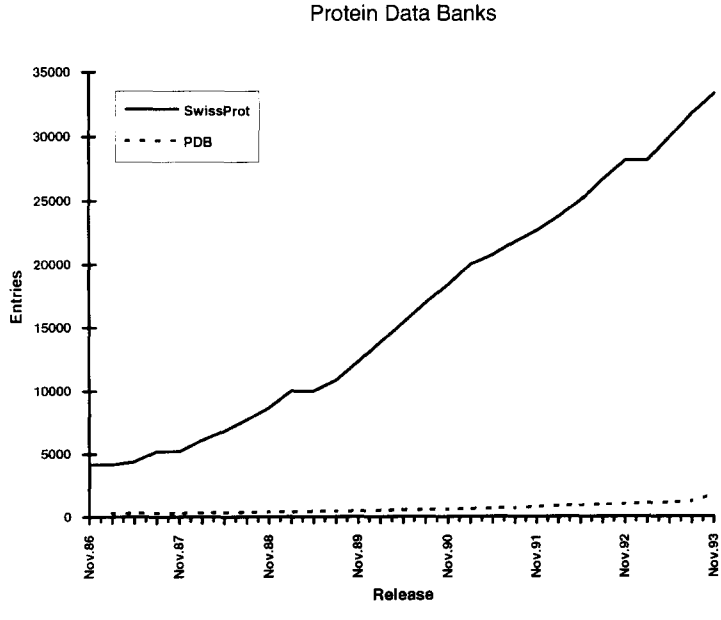


Fig. 3. The sequence–structure gap. The number of entries in the SwissProt and Protein Data Bank (PDB) shown as a function of time.

the sequences. The same approach may obviously also be used for sequences that are similar, but not necessarily evolutionary related, and in that case we probably should talk about similarity based modelling. However, in this paper we will use homology based modelling as a general term, especially since the distinction between homology and similarity may be difficult in many cases.

Homology based modelling may turn out to be essential for the future of protein engineering. In Fig. 3, the number of entries in the SwissProt protein sequence database (Bairoch and Boeckmann, 1992) and the Brookhaven protein structure database (Bernstein et al., 1977; Abola et al., 1987) are shown as a function of time. As we can see, there is a very significant gap between the number of sequences and the number of structures. This gap is in fact even larger than shown in Fig. 3, as not all entries in the Brookhaven database are unique structures. A large number of entries are mutants of other structures or identical proteins with different substrates or inhibitors. There has been an exceptional growth in the number of protein structures over the last 2–3 years. However, it is unrealistic to assume that we will be able to get high resolution experimental structures of all known proteins. The structure determination process is too time consuming, and the sequence databases are growing at a far faster pace, as shown in Fig. 3, especially as a consequence of several large-scale genome sequencing projects.

On the other hand, it may not really be necessary to do experimental structure determination of all proteins (Ring and Cohen, 1993). The assumption that similar sequences have similar structures (see Fig. 4) has been proved valid several times and it seems to be true even for short peptide sequences as long as they come from proteins within the same general folding class (Cohen et al., 1993). An interesting case which is to some degree an exception to this rule is the structure of HIV-1 reverse transcriptase (Kohlstaedt et al., 1992). Two units with identical sequence have similar secondary structure, but very different tertiary structure. However, this seems to be a rather exceptional case. New approaches to general structure alignment (Orengo

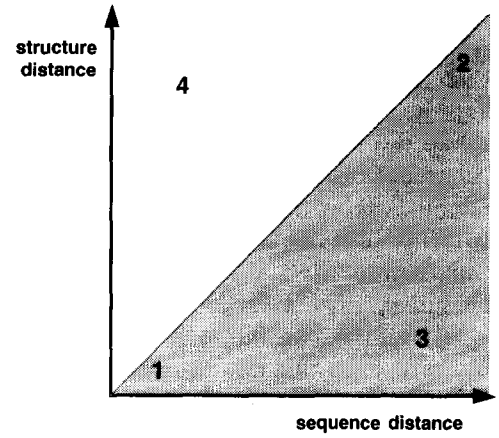


Fig. 4. Sequence and structure similarity. In most cases similar sequences have similar structures (region 1), and dissimilar sequences (i.e., measured by a standard mutation matrix) have dissimilar structures (region 2). In several cases quite dissimilar sequences have been shown to have very similar structures, at least with respect to individual domains (region 3). In very special cases we may have similar sequences with different structures (region 4), at least with respect to tertiary structure, showing that environment and binding to other proteins may be essential for the final conformation in some cases. However, in most cases it seems to be safe to assume that structures can be found in the lower grey triangle of this graph, indicating that structure is better conserved than sequence.

et al., 1992; Holm et al., 1992; Alexandrov and Go, 1993; Lessel and Schomburg, 1993) have made it possible to search for structurally conserved domains in proteins with very low sequence similarity (Swindells, 1994). This is an important approach, as structure normally is better conserved than sequence (Doolittle, 1992). Several cases have been identified where the sequences are very different (at least by traditional similarity measures), whereas the three-dimensional structures are surprisingly similar. The identification of a globin fold in a bacterial toxin (Holm and Sander, 1993), and the similarity between the DsbA protein and thioredoxin (Martin et al., 1993) are relevant examples. Recently the structure of the human serum amyloid P component was shown to be similar to concanavalin A and pea lectin, despite only 11% sequence identity (Emsley et al., 1994), and the similarity between hen egg-white lysozyme and a lysozyme-like domain in bacterial muramidase "is remarkable

in view of the absence of any significant sequence homology", as noted by Thunnissen et al. (1994). This shows that there probably is a limited number of protein folds, and this number must be lower than the number of sequence classes, defined as groups of similar protein sequences. Recent estimates show that this number probably is close to 1000 different protein folds (Chothia, 1992), and approx. 160 of these folds are known so far (Burley, 1994; Orengo et al., 1993). This means that rather than full structure determination of a very large number of proteins, it may be sufficient to do structure determination of only a few selected examples of each protein fold, and use this as a basis for homology based modelling of other proteins shown to have the same fold.

Homology based modelling of the 3-D structure of a novel sequence can be divided into several steps. First, one or more templates must be identified, defined as known protein structures assumed to have the same fold as the trial sequence. Then a sequence alignment between trial sequence and template is defined, and based upon this alignment an initial trial model can be built. This initial model must be refined in several steps, taking care of gap splicing, loops, side chain packing etc. The final model can be evaluated by several quality criteria for protein structures. An example of homology based modelling is the modelling of cinnamyl alcohol dehydrogenase based on the structure of alcohol dehydrogenase (McKie et al., 1993).

4.1. Identification of folding class

The protein folding problem is a fundamental problem in structural biology. This problem can be defined as the ab initio computation of a protein's tertiary structure starting from the protein sequence. This problem has not been solved and appears to be extremely difficult. If we want to solve the problem by computing an energy term for all conformations of a protein, defined by rotation around the ϕ and ψ backbone angles of N residues in 10 degree steps, we have to evaluate $36^{2(N-1)}$ alternatives, even without considering the side chains. For a peptide with 15 residues this corresponds to 10^{44} conformations. A hypothetical computer with 10^6 processors, each processor running at 10^{15} Hz (the frequency of UV light) and completing the energy evaluation of one conformation per cycle would need 3×10^{15} years in order to test all conformations. The estimated age of the universe is 14×10^{12} years. A more realistic approach is the use of molecular dynamics or Monte Carlo methods for simulation of protein folding. However, it is still very difficult to use this as an ab initio approach, both because folding is a very slow process compared to a realistic simulation time scale, and also because it is very difficult to distinguish between correctly and incorrectly folded structures using standard molecular mechanics force fields (Novotny et al., 1984). A possible alternative approach may be to generate potential folds on a simplified lattice representation of possible residue positions (Covell and Jernigan, 1990; Crippen, 1991). However, this approach is still very experimental.

Some progress has been achieved in the area of secondary (rather than tertiary) structure prediction (Benner and Gerloff, 1993). Studies of local information content indicate that 65% match may be an upper limit for single-sequence prediction methods (Rao et al., 1993), whereas methods taking homology data into account may probably raise this limit to approx. 85%. Methods based on neural networks and combinations of several prediction schemes seem to give good predictions, and especially methods using homology data from multiple alignments may give predictions at 70% match or better in many cases (Salzberg and Cost, 1992; Boscott et al., 1993; Rost and Sander, 1993a; Rost et al., 1993; Levin et al., 1993). Also methods taking potential residue–residue interactions into account, like the hydrophobic cluster analysis (HCA), may be used for identification of potential secondary structure elements (Woodcock et al., 1992). It has been shown that by restricting the prediction to a consensus region with stable conformation it is possible to make very reliable predictions (Rooman and Wodak, 1992). In one case, neural networks were shown to be capable of returning a limited amount of information on the tertiary structure (Bohr et al., 1993).

The prediction methods depend upon a training set of known structures classified into secondary structure elements by an automated assignment method, like DSSP (Kabsch and Sander, 1983), Pcurve (Sklenar et al., 1989) or Define (Richards and Kundrot, 1988). It is a potential problem that the automatic assignment of secondary structure types to known structures may

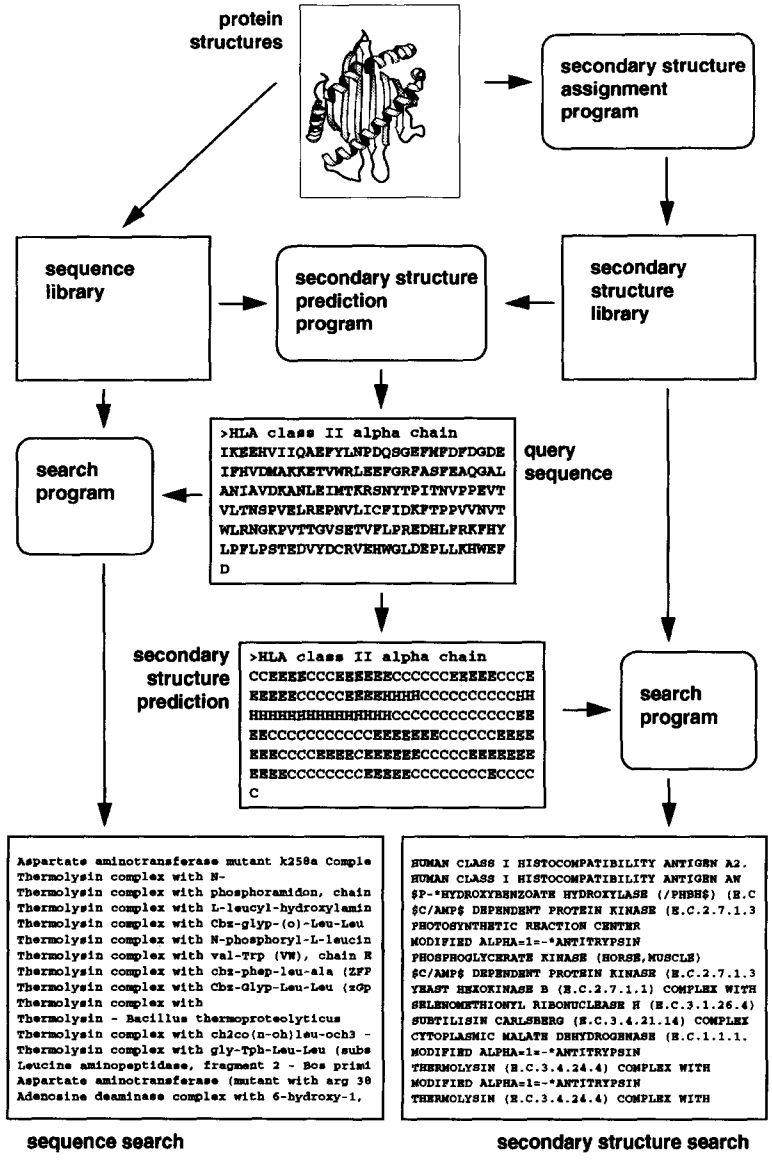


Fig. 5. Structure retrieval by secondary structure. A flow chart for structure retrieval by secondary structure (right side) compared to retrieval by sequence (left side). Please see the text for details. In this example the Secondary Structure Library was generated using DSSP (Kabsch and Sander, 1983), the secondary structure was predicted with the PHD program (Rost et al., 1994), and Fasta (Pearson, 1990; Pearson and Lipman, 1988) was used to search the Secondary Structure Library and the NRL-3-D databases (Namboodiri et al., 1988; George et al., 1986). Only the secondary structure based method was able to identify the HLA class I structure as similar to the class II structure. The ribbon representation of the HLA class I antigen binding region used in this figure was generated with Molscript (Kraulis, 1991).

be inconsistent, compared to the more sophisticated classification which can be achieved by a trained expert. Recent studies show that the average agreement between alternative assignment methods used on identical structures is close to 65% for three methods (Colloc'h et al., 1993), or 79% if only two methods are compared (Woodcock et al., 1992). Vadar is a new classification method which is aiming at a better agreement between manual and automatic assignment (Wishart et al., 1994), to what degree this may have influence on prediction systems remains to be seen.

Over the last few years it has been realised that the *inverse* folding problem is much easier to solve (Bowie et al., 1991; Blundell and Johnson, 1993; Bowie and Eisenberg, 1993). The inverse folding problem can be defined as follows: given a known protein structure, identify all protein sequences which can be assumed to fold in the same way. A large number of protein structures must be available in order to use this as a general approach, as the relevant protein fold has to be represented in the database in order to be identified. However, with a limited number of possible folds actually used by Nature, a complete database of all folds appears to be possible.

Some information about possible folding classes can be derived from experimental data. Circular dichroism can be used as a crude way of measuring the relative amounts of secondary structure in a protein. Classification methods based on amino acid composition can be used for classification of proteins into broad structural classes (Zhang and Chou, 1992; Zhou et al., 1992; Chou and Zhang, 1992; Dubchak et al., 1993). This information may limit the number of different folds which have to be evaluated. It is also possible that such information may be used to improve the performance of other methods, although data on secondary structure prediction of all-helical proteins seems to indicate that the gain may be small (Rost and Sander, 1993b). However, for a unique identification of folding class more sensitive methods are needed, and the most useful one is probably some kind of protein sequence library search.

In order to identify the folding class we have to search a database of known protein structures with our trial sequence. The problem is that standard methods for sequence retrieval may not be sensitive enough in all cases. If the sequences are similar, then retrieval is trivial. However, we know that there are cases where structures are known to be similar despite very different sequences. How can these cases be identified in a reliable way?

The most promising approaches are based on methods for describing the environment of each residue (Bowie et al., 1991; Eisenberg et al., 1992; Overington et al., 1992; Ouzounis et al., 1993; Wilmanns and Eisenberg, 1993; Lüthy et al., 1994). This description can be used for generating a profile, showing to what degree each residue is found in a similar environment in other structures, and this profile can be used as a basis for sequence alignment and library searches. Similar property profiles can also be used for searching database systems of protein structures (Vriend et al., 1994).

A very simple approach can be used if we accept the hypothesis that protein sequences representing structures with a similar linear distribution of secondary structure elements may fold in a similar way. We can then create a sequence type library of known structures where the residues are coded by secondary structure codes rather than residue codes (see Fig. 5). Given the sequence of a protein with unknown 3-D structure, we can use a secondary structure prediction method and translate the sequence into a secondary structure description. If we define a suitable 'mutation' matrix describing the probability of inter conversion between different secondary structure elements, then a standard library search program like Fasta (Pearson and Lipman, 1988; Pearson, 1990) can be used in order to identify potential template structures. The example shown in Fig. 5 is the identification of HLA class I as a suitable candidate for homology modelling of HLA class II. The sequence similarity is very low, 11% sequence identity in the antigen binding region (based on alignment of the structures), and especially for this region most sequence based methods will retrieve a large number of alternative sequences before any of the class I molecules.

However, for the secondary structure based approach the HLA class I sequences are retrieved as top candidates. The structure prediction did not include any information about the HLA class II structure, which recently has been published (Brown et al., 1993). It should be mentioned that the 11% sequence identity score is not significantly higher than the score from a random alignment of sequences. If we, for each sequence in the SwissProt protein sequence library, align it against a sequence selected at random from the same library (alignment without gaps, using the full length of the shortest sequence, and start the alignment at a random position within the longest sequence), then the average percentage of identical residues is $(6 \pm 6)\%$ at 3 standard deviations.

The identification method using secondary structure is based on an assumption which has to be examined more closely, and the implementation of it is very crude. Much work can be done on the secondary structure prediction, the 'mutation' matrix and the search method. It will probably improve the performance to use a position-dependent gap penalty, where most gaps are placed in loop regions rather than in helices or strands. However, the method is very simple to implement and test, as necessary tools and data already are available in most labs.

4.2. Sequence alignment

As described in the introduction a crucial feature in molecular evolution has been the parallel exploration of several different mutations. And although mechanisms like horizontal gene transfer and intragenic recombination may have been important as key steps in the evolution of new proteins, the most common mechanism seems to have been gene duplication followed by mutational modification (Doolittle, 1992). This means that especially multiple sequence alignment can give essential information about the mutation studies already performed by Nature. Conserved residues are normally conserved because they

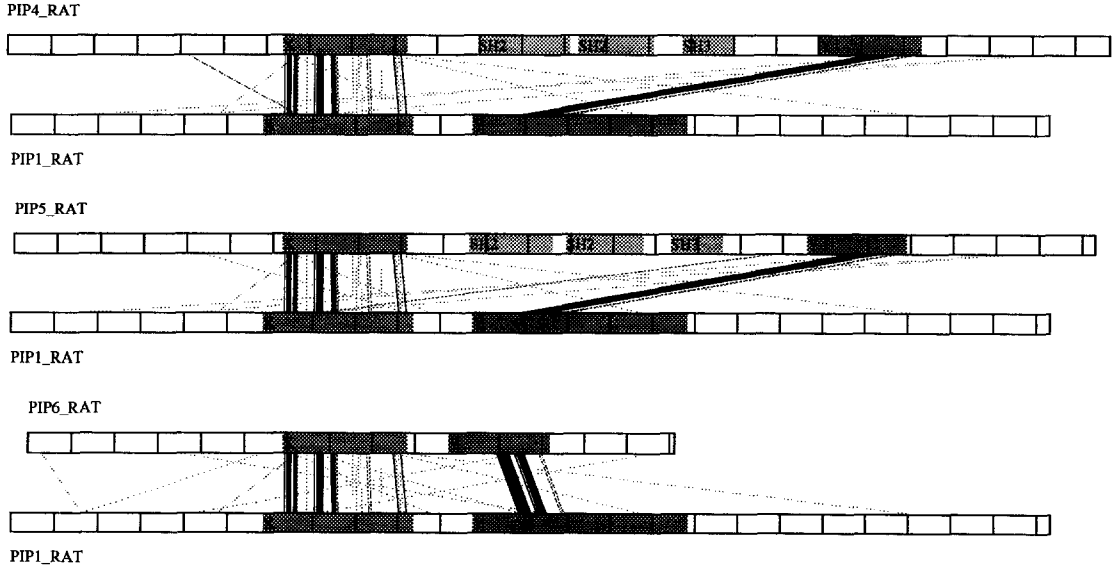


Fig. 6. Multim alignment. Alignment of inositol triphosphate specific phospholipase C β 1 from rat (PIP1_RAT) against three other PIP sequences from rat. Each horizontal bar represents a sequence, marked in 50 residue intervals. Black lines connecting the bars represent well conserved motifs found in all sequences, in this case subsequences of 8 residues where at least 4 residues are completely conserved in all 4 sequences. It is very easy to identify two well conserved regions, annotated as region X and Y in the SwissProt entries, despite a 400 residue insertion in two of the sequences. This insertion contains SH2 and SH3 domains (Pawson, 1992). It is an interesting observation that the extra C-terminal domain of the PIP1_RAT sequence shows a weak similarity to myosin and tropomyosin sequences.

have an important structural or functional role in the protein, and identification of such residues will thus give vital information about structure and activity of a protein.

Several tools have been developed for multiple alignment. A very attractive one is Macaw (Schuler et al., 1991), which will generate several alternative alignments of a given set of regions, and in a very visual way help the user to identify a reasonable combination of (sub)alignments.

An even more general tool is Multim (Drabløs and Petersen, 1994). Here all possible alignments, based on short motifs, are shown simultaneously, and the user is free to identify potential similarities even in cases with low sequence identity and very disperse motifs. This is possible because of the superior classification potential of the human brain compared to most automatic approaches. The method includes an option for probability based filtering of motifs, and an example of a Multim alignment is shown in Fig. 6.

However, it is important to realise that in standard sequence alignment we are trying to solve a three-dimensional problem (residue interactions) by using an essentially one-dimensional method (alignment of linear protein sequences). As a consequence important conserved through–space interactions may not be evident from a standard sequence alignment. A good example can be found in the alignment of lipases (Schrag et al., 1992). In Fig. 7, the sequence alignment of residues in a structurally conserved core of three lipases (*Rhizomucor miehei* lipase (Derewenda et al., 1992), *Candida antarctica* B lipase (A. Jones, personal communication) and human pancreatic lipase (Winkler et al., 1990) is shown. The active site residues, Ser (S), Asp (D) and His (H), are shown as black boxes. The Ser and His residues are at identical positions. However, the Asp residue of the pancreas lipase is at a very different sequence position compared to the other two lipases. It would be very difficult to identify this as the active site Asp from a sequence alignment. If we look at the structural alignment in Fig. 8, we see that the positions are structurally equivalent, it is possible for all three lipases to have highly similar relative orientation of the active site atoms, despite the fact that the alternative Asp positions are located at the end of two different β-strands.

An improved alignment may be generated if we can incorporate 3-D data for at least one of the sequences in the linear alignment (Gracy et al., 1993). However, in order to get a reliable alignment of sequences with low sequence similarity, we have to take true three-dimensional effects into account. This means that if we are able to identify a known 3-D structure as a potential basis for modelling, then the sequence alignment should be done in 3-D using this structure as a template. This can be done by threading the sequence through the structure and calculating pairwise interactions (Jones et al., 1992; Bryant and Lawrence, 1993).

4.3. Model refinement

As soon as a template has been identified, and an alignment between this template and a sequence has been defined, a 3-D model of the protein can be generated. We can either use the template coordinates directly, combined with dif-

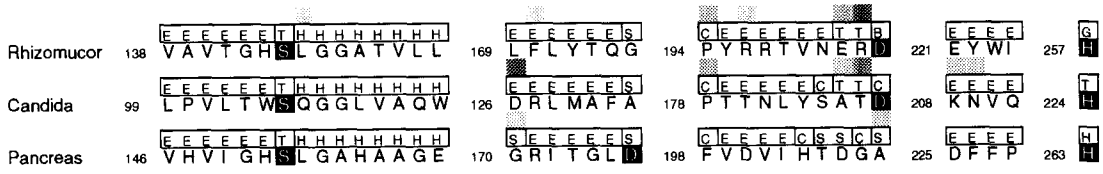


Fig. 7. Sequence alignment of lipases. Alignment of structurally conserved regions of three lipases. For each lipase the solvent accessible surface in % compared to the GXG standard state (grey scale, white is buried and black is exposed), the secondary structure as defined by the DSSP program, and the sequence is shown. The position of each subsequence in the full sequence is also shown. The active site residues are shown in white on black. Please observe the shift of the active site Asp (D) between two very different positions. The alignment was generated using Alscript (Barton, 1993).

ferent modelling approaches for the ill-defined regions, or the template can be used as a more general basis for folding the protein by distance geometry (Srinivasan et al., 1993) or general molecular dynamics methods. Loop regions are often highly variable, and must be treated with special approaches (Topham et al., 1993).

It is also necessary to consider the orientation of side chains. Although the backbone may be well conserved, many residues especially at the protein surface will be mutated, as shown in Fig. 9. The stability of a protein depends upon an optimal packing of residues, and it is important to optimise side chain conformation if we want to study protein stability and complex formation. A very common approach is the use of rotamer libraries combined with molecular dynamics refinement. Recent studies show that this step of the modelling in fact may be less difficult than has been assumed (Eisenmenger et al., 1993).

4.4. Model evaluation

A protein model based on homology (or similarity) has to be verified in as many ways as possible, and experimental methods should always be preferred. Mutation studies may give valuable information about active site residues and important interactions, and exposed regions may to some degree be identified by using antibodies. However, in many cases the rational for modelling by homology is the very lack of experimental data related to structure, and we have to use other more general methods for evaluation of models.

Some of the approaches we already have described for sequence alignment can obviously also be used for evaluation of models. In general, model evaluation can be based on 3-D profiles (Lüthy et al., 1992), contact profiles (Ouzounis et al., 1993) or more general energy potentials (Hendlich et al., 1990; Jones et al., 1992; Nishikawa and Matsuo, 1993). Some of these approaches have been implemented as programs for evaluation of structures or models, like ProCheck (Laskowski et al., 1993) and Prosa (Sippl, 1993a, b). However, in general no model (or even experimental structure) should be trusted beyond what can be verified by experimental methods.

5. NMR of proteins

A prerequisite for rational protein engineering is 3-D structure information about the protein. In

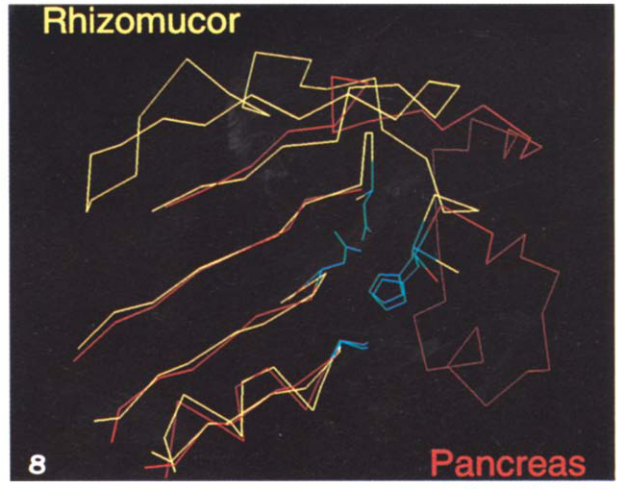


Fig. 8. Structure alignment of lipases. Structure alignment of two of the sequences shown in Fig. 6, including parts of the sequences connecting the core regions. The active site Asp is able to maintain a similar relative orientation, despite very different sequence positions. The alignment was generated using *Insight* (Biosym Technologies).

adddition to X-ray crystallography, NMR is the most important method for protein structure determination.

X-ray crystallography has several advantages when compared to NMR. Solving the crystal structure by X-ray crystallography is usually fast as soon as good crystals of the protein are obtained (even if it may not be so easy to obtain these crystals). It is also possible to determine the structure of very big proteins. The major disadvantage of X-ray crystallography is that it is the crystal structure that is determined. This implies that crystal contacts may distort the structure (Chazin et al., 1988; Wagner et al., 1987). Since active sites and other binding sites usually are located on the surface of the proteins, very important regions of the protein may be distorted. Some structures even show large differences between NMR and X-ray structure (Frey et al., 1985; Klevit and Waygood, 1986)

The advantage of NMR is that it is dealing with protein molecules in solution, usually in an environment not too different from its natural one. It is possible to study the protein and the dynamical aspects of its interaction with other molecules like substrates, inhibitors, etc. It is also possible to obtain information about apparent pK_a values, hydrogen exchange rates, hydrogen binding and conformational changes.

5.1. A short introduction to NMR

All nuclei contain protons, and therefore they carry charge. Some nuclei also possess a nuclear spin. This creates a magnetic dipole, and the nuclei will be oriented with respect to an external magnetic field. The most commonly studied nuclei in protein NMR (^{1}H, ^{13}C and ^{15}N) have two possible orientations, representing high and low energy states. The frequency of the transition between the two orientations is proportional to the magnetic field. At a magnetic field of 11.7 Tesla the energy difference corresponds to about 500 MHz for protons. In an undisturbed system there will be an equilibrium population of the possible orientations, with a small difference in spin population between the high and low energy orientation.

The equilibrium population can be perturbed by a radio frequency pulse of a frequency at or close to the transition frequency. In addition, the spins will be brought into phase coherence (concerted motion) and a detectable magnetisation will be created. The intensity of the NMR signal

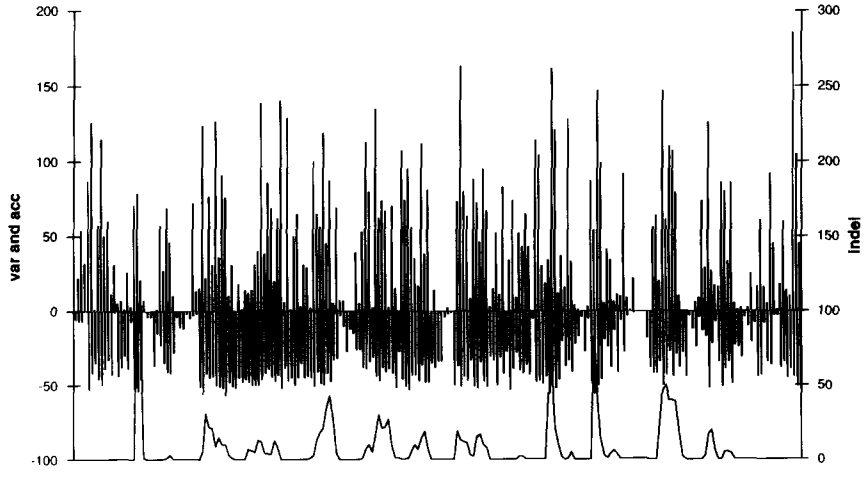


Fig. 9. Sequence variability of trypsin. A graphical representation of the HSSP entry (Sander and Schneider, 1991) of trypsin (1sgt.hsp), showing sequence variability of trypsin from an alignment of 109 sequences. Positive values on the left-hand scale are solvent accessibilities, negative values are sequence variability. Values on the right-hand scale are number of sequences with insertions or deletions at a given position (bottom curve). Buried regions have very low sequence variability and no insertions or deletions.

is proportional to the population difference between the levels the nuclei can possess.

Nuclei of the same type in different chemical and structural environments will experience different magnetic fields due to shielding from electrons. The shielding effect leads to different resonance frequencies for nuclei of the same type. The effect is measured as a difference in resonance frequency (in parts per million, ppm) between the nuclei of interest and a reference substance, and this is called the chemical shift. In molecules with low internal symmetry most atoms will experience different amounts of shielding, the resonance signals will be distributed over a well-defined range, and we get a typical NMR spectrum.

The process that brings the magnetisation back to equilibrium may be divided into two parts, longitudinal and transverse relaxation. The longitudinal or T_1 relaxation describes the time it takes to reach the equilibrium population. The transverse or T_2 relaxation describes the time it takes before the induced phase coherence is lost. For macromolecules the T_2 relaxation is always shorter than the T_1 relaxation. Short T_2 relaxation leads to broad signals because of poor definition of the chemical shift. Most molecules have dipoles with magnetic moment, and the most important cause of relaxation is fluctuation of the magnetic field caused by the brownian motion of molecular dipoles in the solution. How effective a dipole may relax the signal depends upon the size of the magnetic moment, the distance to the dipole, and the frequency distribution of the fluctuating dipoles.

A nucleus may also detect the presence of nearby nuclei (less than three bonds apart), and this will split the NMR signal from the nucleus into more components. Several nuclei in a coupling network is called a spin system.

By applying radio frequency pulses it is possible to create and transfer magnetisation to different nuclei. It is, as an example, possible to create magnetisation at one nucleus, and transfer the magnetisation through bonds to other nuclei where it may be detected. The pulses are applied in a so-called pulse sequence (Ernst, 1992; Kessler et al., 1988).

5.2. Methods for structure determination by NMR

The methodology for determination of protein structure by two-dimensional NMR is described in several textbooks and review papers (Wagner, 1990; Wüthrich, 1986; Wider et al., 1984). The standard method is based on two steps, *sequential assignment*: assignment of resonances from individual amino acids, and *distance information*: assignment of distance correlated peaks between different amino acids.

5.2.1. Assignment of resonances from individual amino acids

The first step involves acquiring coupling correlated spectra (COSY, TOCSY) in deuterium oxide to determine the spin system of correlated resonances. Some amino acids have spin systems that in most cases make them easy to identify (Gly, Ala, Thr, Ile, Val, Leu). The other amino acids have to be grouped into several classes, due to identical spin systems, even though they are chemically different. The spin systems can be correlated to the NH proton by acquiring COSY and TOCSY spectra in water.

The assigned NH resonance is then used in distance correlated spectra (NOESY) to assign correlations to protons (NH, H_α, H_β) at the previous amino acid residue (Fig. 10). By combining the knowledge of the primary sequence (which gives the spin system order) with the NMR data collected it is possible to complete the sequential assignment.

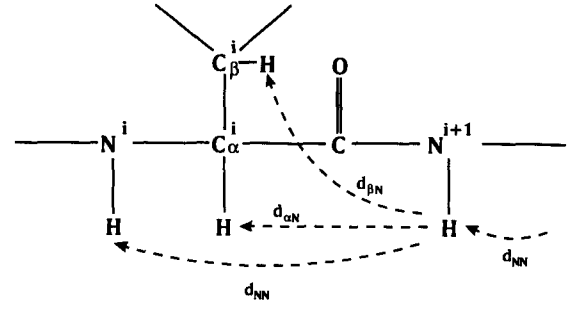


Fig. 10. Short-range interactions in proteins. Part of a peptide backbone with two amino acid residues. Sequential correlations from $NH^{(i+1)}$ to $NH^{(i)}$, $C_\alpha H^{(i)}$ and $C_\alpha H^{(i)}$ are shown.

5.2.2. Assignment of distance correlated peaks

When the sequential assignment is done the assignment of short range nOe (up to four residues) will give information about secondary structure (α-helix, β-strand). Long-range correlations will serve as constraints (together with scalar couplings) to determine the tertiary structure of the protein. Excellent procedures describing these steps are available (Roberts, 1993; Wüthrich, 1986).

With large proteins there will be spectral overlap of resonance lines. The problem is partially solved by labelling the protein with ^{13}C and ^{15}N isotopes. Triple resonance multidimensional NMR methods (Griesinger et al., 1989; Kay et al., 1990) may then be applied. The resonances will then be spread out in two more dimensions (^{13}C and ^{15}N) and the problem with overlap is reduced. These methods depend upon the use of scalar couplings to perform the sequential assignment, the sequential assignment procedure will then be less prone to error. The NOESY spectra of such large proteins are often very crowded, but four-dimensional experiments like the ^{13}C–^{13}C edited NOESY spectrum (Clore et al., 1991b) have been designed. Such experiments will spread the proton–proton distance correlated peaks by the chemical shift of its corresponding ^{13}C neighbour and reduce the spectral overlap. Secondary structure elements may also be predicted from the chemical shift of ^{1}H and ^{13}C (Spera and Bax, 1991; Williamson and Asakura, 1991; Wishart et al., 1992).

5.3. Larger proteins

Obtaining NMR-spectra of proteins has some aspects that should be considered.

Spectral overlap. As we move to larger proteins the probability of overlap of resonance lines increases. At some point it will become impossible to do sequential assignment due to this overlap. Application of 3-D and 4-D multiresonance NMR has made it possible to assign proteins in the 30 kDa range (Foght et al., 1994; Stockman et al., 1992).

Fast relaxation. As the size of the protein is increased the rate of tumbling in solution is reduced. This leads to a reduced transverse relaxation time (T_2), and broadening of the resonance lines in the NMR spectra. The intensities of the peaks are reduced and they may be difficult to detect. The short transverse relaxation time will also limit the length of the pulse sequences it is possible to apply (because there will be no phase coherence left), and multidimensional methods become difficult.

Behaviour of the protein. The proteins for which it is possible to determine a 3-D structure by NMR or X-ray crystallography are probably a subset of all proteins (Wagner, 1993). Proteins may have regions with mobility and few cross peaks. The effective size of a protein is often increased by aggregation. The amount of aggregation can often be reduced by reducing the protein concentration. Thus, very often the degree of aggregation will determine whether it is possible to assign and solve a protein structure by NMR, by limiting the maximum concentration that may be used. The stability of the proteins is also a major issue. A sample may be left in solution for days, often at elevated temperatures, so denaturation may become a problem.

5.4. Other applications of NMR

Photo-CIDNP (chemically induced nuclear polarisation) is an interesting technique for the study of surface positioned aromatic residues in proteins (Broadhurst et al., 1991; Cassels et al., 1978; Hore and Kaptain, 1983; Scheffler et al., 1985). By introducing a dye and exciting it with a laser, it is possible to transfer magnetisation to aromatic residues, where it can be observed.

In addition to high-resolution NMR, solid state NMR has also been applied to studies of proteins. Studies of active sites and conformation of bound inhibitors yields interesting information. The stability of proteins may be monitored under different conditions by detecting signals from transition intermediates bound to the active site (Burke et al., 1992; Gregory et al., 1993). Structural constraints on transition state conformation of bound inhibitors can be obtained (Auger et al., 1993; Christensen and Schaefer, 1993). Structural constraints of the fold and conformation of the

amino sequence may be gathered by setting upper and lower distances for lengths between specific amino acids (McDowell et al., 1993).

Using solid-state NMR it is also possible to study membrane proteins and their orientation with respect to their membrane (Killian et al., 1992; Ulrich et al., 1992). We expect such studies to give insight into ion channels in membranes (Woolley and Wallace, 1992).

5.5. Paramagnetic relaxation

An important mechanism for relaxation in high-resolution NMR is dipolar relaxation. Usually this is induced by the spin of nuclei in the immediate vicinity, and it is a function of the size of the dipole. The electron is also a magnetic dipole, and the magnitude of this dipole is about 700-times that of a proton. Paramagnetic compounds have an electron that will interact with nearby protons and increase the relaxation rate of these protons.

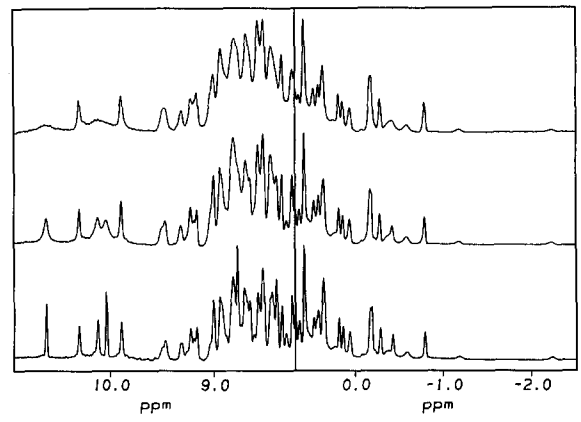


Fig. 12. Lysozyme with Gd-DTPA, 1-D spectra. High (right) and low field region of 1-D ^{1}H-NMR spectra of lysozyme with different concentrations of Gd-DTPA added. The lower trace shows the spectrum of pure lysozyme, the middle and upper trace show the spectra with Gd-DTPA added (middle trace: 5 mM lysozyme/0.25 mM Gd-DTPA, upper trace: 5 mM lysozyme/0.5 mM Gd-DTPA).

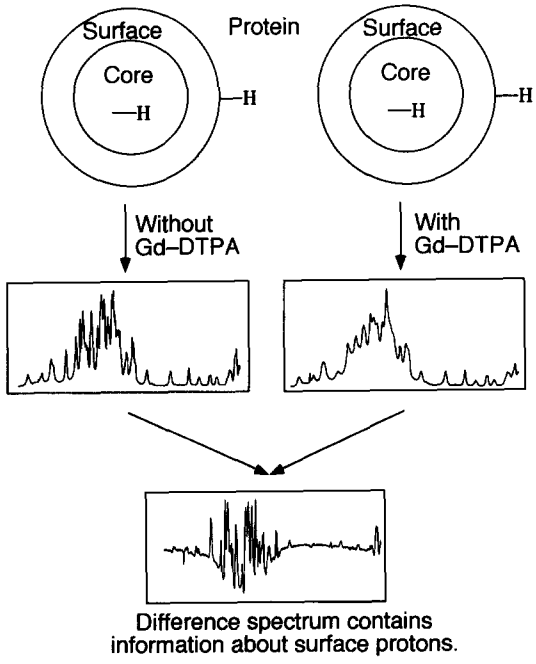


Fig. 11. The paramagnetic relaxation method. Outline of the paramagnetic relaxation method. The protons located at the protein surface will be closer to the dissolved paramagnetic relaxation agent than the protons located inside the protein core, hence the resonance lines from protons at the surface will be broadened more than resonance lines stemming from protons located inside the protein.

The widest use of paramagnetic compounds has been of Gd^{3+} bound to specific sites in a protein (Dobson et al., 1978), but also other compounds have been used (Chang et al., 1990; Hernandez et al., 1990a, b). This will make it possible to identify resonance lines from residues in the vicinity of the binding site. It is also possible to calculate distances from the paramagnetic atom as the relaxation effect is distance dependent.

The paramagnetic broadening effect can also be used with a compound moving freely in solution (Drayney and Kingsbury, 1981; Esposito et al., 1992; Petros et al., 1990). In this way residues located on or close to the protein surface will give broadened resonance lines compared to residues in the interior of the protein.

This method can be used to measure important nOe and chemical shifts inside the protein directly, or it can be used as a difference method to identify resonances at the surface by comparing spectra acquired with and without the paramagnetic relaxation agent (Fig. 11).

We have used the paramagnetic compound gadolinium diethylenetriamine pentaacetic acid (Gd-DTPA) as a relaxation agent. Gd-DTPA will

increase both the longitudinal and the transverse relaxation rates of protons within the influence sphere. Suitable NMR experiments to highlight the relaxation effect may be NOESY, ROESY and TOCSY (Bax and Davis, 1985; Braunschweiler and Ernst, 1983)

Gd-DTPA is widely used in magnetic resonance imaging (MRI) to enhance tissue contrast. It is assumed to be non-toxic and we do not expect it to bind to proteins. We used the well-studied protein hen egg-white lysozyme as a test protein. Both the structure and the NMR spectra of this protein are known (Diamond, 1974; Redfield and Dobson, 1988), and the protein is extremely well suited for NMR experiments.

In Fig. 12, the 1-D ^{1}H-NMR spectrum recorded in the presence and absence of Gd-DTPA is shown. Although it is evident that there is a selective broadening in the 1-D spectrum, it is also clear that there are problems with overlapping spectral lines. We therefore applied two-dimensional NMR methods, and shown in Fig. 13 is the low field region of a NOESY spectrum of lysozyme. The region corresponds to the same region as shown in Fig. 12.

From Fig. 13 we see that the signals from W63, W 63 and W123 disappear with addition of Gd-DTPA, while the signals from W28, W108 and W111 still are observable. By examination of the solvent accessible surface of lysozyme it is evident that the indole NH of W62, W63 and W123 is exposed to solvent, while the indole NH of W28, W108 and W111 is not exposed. This shows that the changes in the spectrum are as expected from the structure data.

The appearance of the NH–NH region of the spectrum (Fig. 14) also shows the reduction in the number of signals in the Gd-DTPA exposed spectrum.

This shows that the paramagnetic broadening effect can be used for selective identification of signals from solvent exposed residues in a protein.

6. Modelling of electrostatic interactions

One of the fundamental steps in the protein engineering process shown in Fig. 1 is the design step, where a correlation between structure and properties is established in order to select potential structural candidates that match new functional profiles. The understanding of this correlation implies a realistic modelling of the physical chemical properties involved in the functional features to be engineered. These features are basically of two types: diffusional and catalytic. Any ligand binding to a protein, whether ligand-receptor or substrate-enzyme, is essentially a diffusional encounter of two molecules. Electrostatic interactions are the strongest long-range forces at the molecular scale and, thus, it is not surprising that they are one of the determinant effects in the final part of the encounter (Berg and von Hippel, 1985). In the case of substrate–enzyme interactions the catalytic step that follows the binding of the substrate seems to be possible

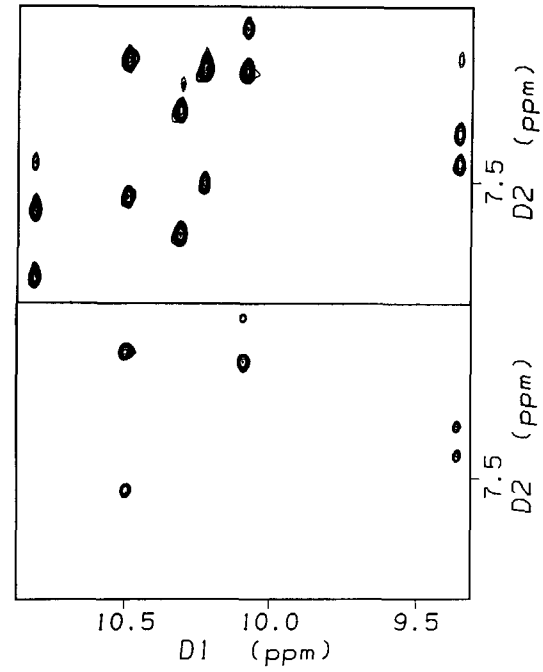


Fig. 13. Lysozyme with Gd-DTPA, NOESY spectra. Low field region of the NOESY spectrum of lysozyme. Correlations between the indole NH of the tryptophan residues with other protons from the same residue is shown. The upper panel shows the normal NOESY spectrum, the lower panel the NOESY spectrum after addition of Gd-DTPA (5 mM lysozyme/0.25 mM Gd-DTPA). The panels are plotted at identical contour levels. The experimental conditions were 5 mM lysozyme, pH 3.8, mixing time 200 ms.

mainly by the presence of electrostatic forces that stabilise the reaction intermediates in the binding site (Warshel et al., 1989), from which the product formation may proceed. Another and much more basic necessary condition for a successfully engineered protein is that a functional folded conformation is maintained. Solvation of charged groups is one of the determinants in protein folding (Dill, 1990), so that even the conformation of the protein is electrostatically driven. Given the ubiquitous role of electrostatic interactions, it is then obvious that their accurate modelling is an essential prerequisite in the design of engineered proteins.

Several good reviews exist on protein electrostatics (Warshel and Russel, 1984; Matthew, 1985; Rogers, 1986; Harvey, 1989; Davies and McCammon, 1990; Sharp and Honig, 1990). This section intends to give a brief overview of the subject. We start by presenting the methods one can use to model electrostatic interactions. The most familiar methodology in biomolecular modelling is certainly molecular mechanics (MM) (either through energy minimisations or molecular dynamics (MD)). We point out some of the limitations of MM in the treatment of electrostatic interactions, and the need to use alternative ways of describing the system, such as continuum methods. The computation of pH-dependent properties and some potential extensions of MM are also discussed. Finally, we refer some applications of electrostatic methods relevant to protein engineering.

6.1. Molecular mechanics (MM)

In MM simulations, electrostatic interactions are usually described with a pairwise coulombic term of the form q_1q_2/Dr_{12},were q_1 and q_2 are the charges of the pair of atoms, r_{12} their distance, and D the dielectric constant. D is usually set equal to 1 when the solvent is included. A complete simulation in a sufficiently big box with water molecules should, in principle, give a realistic description of the protein molecule (Harvey, 1989). This would be specially true if a force field including electronic polarizability effects (see 6.3.) was available for use with biomolecular systems, which unfortunately is not the case (Harvey, 1989; Davis and McCammon, 1990). We use the term force field in this context as including both the functional form and parameters describing the energetics of the system, from which the forces are derived.

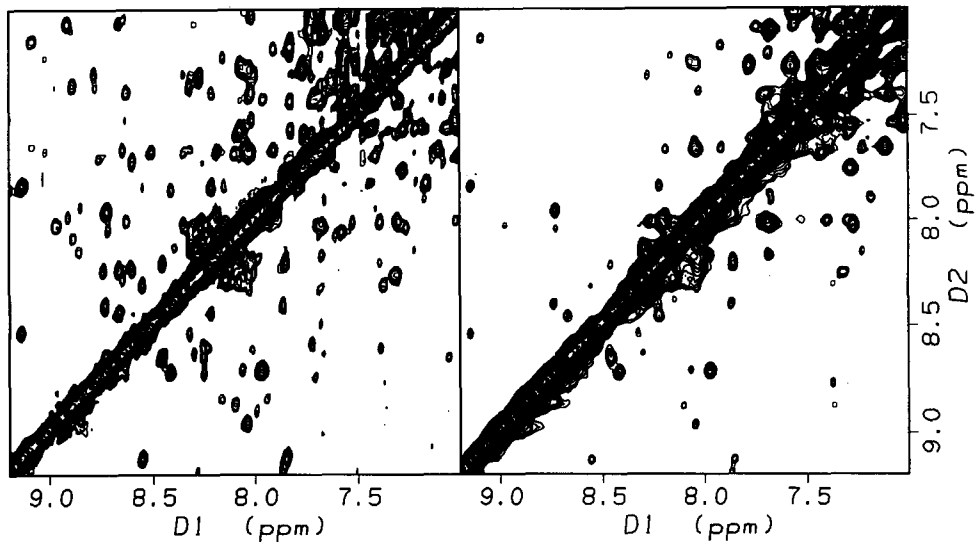


Fig. 14. Lysozyme with Gd-DTPA, NOESY spectra, NH region. NH–NH region of the NOESY spectrum of lysozyme. The normal spectrum is plotted in the left panel, the spectrum acquired with Gd-DTPA in the right panel. The experimental conditions were the same as in Fig. 13.

Simulations where solvent molecules are not treated explicitly are naturally appealing, since the computation time increases with the square of the number of atoms. Several methods have been proposed that attempt to account for solvent effects. The more popular approach is an ad hoc dielectric 'constant' proportional to the distance (e.g., McCammon and Harvey, 1987) but different distance dependencies can be used (e.g., Solmajer and Mehler, 1991). A variety of more elaborated methods were also suggested (Northrup et al., 1981; Still et al., 1990; Gilson and Honig, 1991). All these methods should be viewed as attempts of including solvent screening effects in a simplified way. They can be useful when inclusion of water is computationally prohibitive, but they cannot substitute for an explicit inclusion of solvent since, e.g., the existence of hydrogen bonding with the solvent is not properly described by these approaches.

MM of biomolecules has, in general, heavy computation needs. The number of water molecules that should be included in order to simulate a typical protein in a realistic way is quite large, especially if one wants to perform MD. Also, each pair of atoms has its own electrostatic interaction and the number of pairs cannot be lowered by a short cut-off distance (e.g., 7 Å) as in van der Waals interactions, since electrostatic interactions are very long range, typically up to 10 Å.

MM simulations have also some limitations on the description of the system, since pH and ionic strength effects usually are difficult or impossible to include. The only way to include pH effects is through the protonation state of the residues. Each titrable group (in Asp, Glu, His, Tyr, Lys, Arg, C- and N-terminal) in the protein have two states, protonated or unprotonated. Thus, a protein with N titrable groups will have 2^N possible protonation charge sets. The best we can do is to choose the set corresponding to the protonation states of model compounds at the desired pH. Free ions can be included in MD simulations of proteins (Levitt, 1989; Mark et al., 1991), but it is not clear if the simulated time intervals are long enough to realistically reflect ionic strength effects.

Another problem with MM is that the understanding it provides of the system (through energy minimisation or MD) does not include entropic aspects explicitly, i.e., it does not give free energies directly. There are methods to calculate free energies based on MM potentials (Beveridge and Dicapua, 1989), but even though several applications have been made on biomolecular systems (for a review see Beveridge and Dicapua, 1989), they are still too demanding for routine use in systems of this size. Then, when the properties under study are related to free energies rather than energies (which is often the case), MM by itself can only be seen as a first approach.

In summary, although MM simulations can provide some unique information on the structural and dynamical behaviour of biomolecular systems, some limitations exist due to both conceptual and practical reasons, in particular regarding the treatment of electrostatic interactions. Fortunately, other methods exist that can provide insight on aspects whose modelling is poor or absent in MM simulations, although at the cost of the atomic detail in the description. There is no 'best' modelling method and we should resort to the several methods available in order to gain an understanding of the system that is as complete as possible.

6.2. Electrostatic continuum models

The so-called continuum or macroscopic models assume that electrostatic laws are valid at the protein molecular level and that macroscopic concepts such as dielectric properties are applicable. Protein and solvent are treated as dielectric materials where charges are located. These charges may be titrable groups (whose protonation state may vary), permanent ions (structural and bound ions, etc.) or, more recently, permanent partial charges of polar groups. Given the dielectric description of the system and the placement of the charges, the problem can be reduced to the solution of the Poisson equation (or any equivalent formulation), as can any problem of electrostatics (e.g., Jackson, 1975). The electrostatic potential thus obtained can be used to

study diffusional processes or visually compare different molecules (see 6.6.).

The simplest continuum model assumes the same dielectric constant inside and outside the protein. Typically, a value somewhere between the protein and solvent dielectric constants has been used (Sheridan and Allen, 1980; Koppenol and Margoliash, 1982; Hol, 1985). This approach completely ignores the effects of having two very different dielectric regions, but can be used for a first qualitative computation.

The more common continuum models treat the protein as a low dielectric cavity immersed in a high dielectric medium, the solvent. The way the charges are placed in this cavity and the way the electrostatic problem is solved vary with the particular method. Analytical solutions can be obtained for the simplest shapes, such as spheres, but in general the more complex shapes require numerical techniques.

In the first cavity model the protein was assumed to be a sphere with the charge uniformly distributed over its surface (Linderstrøm-Lang, 1924). Tanford and Kirkwood (1957) proposed a more detailed model in which each charge has a fixed position below the surface. Assuming a spherical geometry allows for a simple solution to the electrostatic problem. It is even possible to include an ionic atmosphere that accounts for ionic strength effects (leading to the Poisson-Boltzman equation). The effect of pH occurs naturally in the formalism. The energy cost of burying a charge inside the low-dielectric protein (self-energy) is taken to be the same as in small model compounds, since at the time when this method was developed (before protein crystallography) charges were believed to be restricted to the protein surface. This limits the method to proteins without buried charges, unless we have some estimate on the self-energy. There are, obviously, some problems in fitting real, irregular-shaped proteins to a spherical model. Some solutions to this problem were proposed, including an ad hoc scaling of interactions based on solvent accessibility (Shire et al., 1974), and the placing of more exposed charges in the solvent region (States and Karplus, 1987).

The inclusion of non-spherical geometries implies the use of numerical techniques, as referred above. Warwicker and Watson (1982) and Gilson et al. (1987) used the finite differences technique to solve, respectively, the Poisson and Poisson-Boltzman equations. Self-energies can be included (Gilson and Honig, 1988), such that the method is fully applicable when buried charges exist. The intrinsic discretization of the system in the finite differences technique, makes these methods readily applicable to any kind of spatial dependency on any of the properties involved. The inclusion of a spatially-dependent dielectric constant, for instance, will be relatively simple. Other extensions such as additional dielectric regions (ligands, membranes, etc.), eventually with charges, should also be possible. Alternative numerical techniques for solving the Poisson or Poisson-Boltzman equations have also been used, including finite elements (Orttung, 1977) and boundary elements (Zauhar and Morgan, 1985).

6.3. Inducible dipole model

The dielectric constant in a region comes from the existence of dipoles in that region, permanent or induced. Permanent dipoles are due to atomic partial charges (e.g., water dipole, peptide bond dipole). Induced dipoles are due to the polarizability of electron clouds. Warshel and Levitt (1976) represented this electronic polarizability by using point dipoles in the atoms. As pointed out by Davies and McCammon (1990) this representation is roughly equivalent to a spatially-dependent dielectric constant. This approach is usually combined with a simplified representation of water by a grid of dipoles (Warshel and Russel, 1984). Ionic strength and pH effects are not considered.

6.4. pH dependency

All the above methods deal with a particular charge set (see 6.1.), even when pH effects are considered. However, a protein in solution does not exist in a single charge set. We are usually interested in the properties of a protein at a given pH and ionic strength, not at a particular charge set. Moreover, if we want to test the available

methods, we have to test them against experimental results which usually do not correspond to a specific charge set. A common test on the accuracy of electrostatic models is their ability in predicting pK_a values of titrable groups in a protein (see 6.6.), obtained via titrations, NMR, etc. These values can be quite different from the ones of model compounds, due to environment of the groups in the protein. This difference (pK_a shift) can be of several pK units. The experimentally determined apparent pK_a (pK_{app}) is determined as the pH value at which half of the groups of that residue are protonated in the protein solution, i.e., when its mean charge is 1/2 (thus, the equivalent notation $pK_{1/2}$). Then, if we can devise a method to compute the mean charge of the titrable groups at several pH values, we can predict their pK_{app} values.

As mentioned above (see 6.1.), we have 2^N possible charge sets. Any structural property can, in principle, be computed through a Boltzman sum over all those sets, with each one contributing according to its free energy (taken as the electrostatic energy) (Tanford and Kirkwood, 1957; Bashford and Karplus, 1990). The property thus computed is characteristic of the chosen pH value (and ionic strength, if considered) instead of a specific charge set. We are particularly interested in computing the mean charges at a given pH (see last paragraph). A sum with 2^N terms is not, however, a trivial calculation in terms of computer time. Tanford and Roxby (1972) avoided the Boltzman sum by placing the mean charges directly on the titrable groups, instead of using one of the integer sets. This corresponds to considering the titration of the different groups as independent (a mean field approximation; Bashford and Karplus, 1991). Other alternatives to the Boltzman sum are the Monte Carlo method (Beroza et al., 1991), less drastic mean field approximations (Yang et al., 1993; Gilson, 1993), the 'reduced site' approximation (Bashford and Karplus, 1991), or even assume that the predominant charge set is enough to describe the system (Gilson, 1993).

Since electrostatic interactions in proteins are typically dominated by titrable groups whose charge is affected by pH, no electrostatic treatment can be complete without taking this effect into account. A simple, although effective, way of doing this is to: (i) compute the electrostatic free energies (e.g., by a continuum method); (ii) compute the mean charge of each titrable group at a given pH (e.g., by a mean field approximation); (iii) use those charges to compute the electrostatic potential (e.g., by a continuum method), which can be displayed together with the protein structure (see the human pancreatic lipase example in section 6.6.). In this way a pH-dependent electrostatic model of the protein can be obtained, which is not possible with usual MM-based modelling techniques.

6.5. Molecular mechanics revisited

As stated above (see 6.1.), electronic polarizability is not explicitly considered in common force fields. Van Belle et al. (1987) included the induced dipole formalism (Warshel and Levitt, 1976) in MM calculations. The electrostatic interactions in the applied force field were simply 'corrected' with additional terms due to inducible dipoles. However, it should be noted that a force field fitted to experimental data without polarizability terms, should be fitted again if those terms are included.

The protein conformation used in molecular modelling is usually an experimentally based (X-ray, NMR) mean conformation, characteristic of those particular experimental conditions. That conformation may, however, be inadequate for modelling the protein properties at different conditions. In particular, proteins are known to denaturate at extreme pH conditions. Thus, pH-dependent methods such as the continuum methods may give incorrect results when using one single conformation over the whole pH range. Actually, MD simulations have shown that the results can be highly dependent on side chain conformation (Wendoloski and Matthew, 1989). Although overall properties like titration curves did not seem to be very sensitive, individual pK_a's showed variations up to 2.0 pK units.

As mentioned in section 6.1, MM has the problem of what charge set to use in simulations. Instead of using a charge set corresponding to

model compounds at the intended pH, one may use the predominant charge set of the protein, determined, e.g., by a continuum method, as suggested by Gilson (1993).

A different approach to this problem would be to devise a way of including the averaged effect of all charge sets in the MM simulation. We have recently developed a method where a force field is derived which includes the proper averaged effect of all charge sets (a potential of mean force) (to be published). The method depends on the calculation of electrostatic free energies obtained from, e.g., a continuum method.

6.6. Applications

The electrostatic potential, computed in some of the referred methods, can help to understand the contribution of electrostatic interactions in the diffusional encounters of proteins with ligands (substrates or not). The diffusional process driven by the electrostatic field can be simulated through Brownian dynamics (BD) and diffusion rates may be computed (for references see, e.g., Davies and McCammon, 1990). The effect of mutations on the diffusion of superoxide ion into the active site of superoxide dismutase has been studied by this technique (Sines et al., 1990) and faster mutants showing 2–3-fold increase in reaction rate could be designed (Getzoff et al., 1992), although this enzyme usually is considered to be 'perfect'. Electrostatically driven BD simulations can help to reveal steric 'bottlenecks' (Reynolds et al., 1990) and orientational effects (Luty et al., 1993). This method can also be applied to study the encounter of two proteins (Northrup et al., 1988).

Visual comparison of electrostatic fields can also provide useful information. Soman et al. (1989) showed that rat and cow trypsins have similar electrostatic potentials near the active site, despite a total charge difference of 12.5 units.

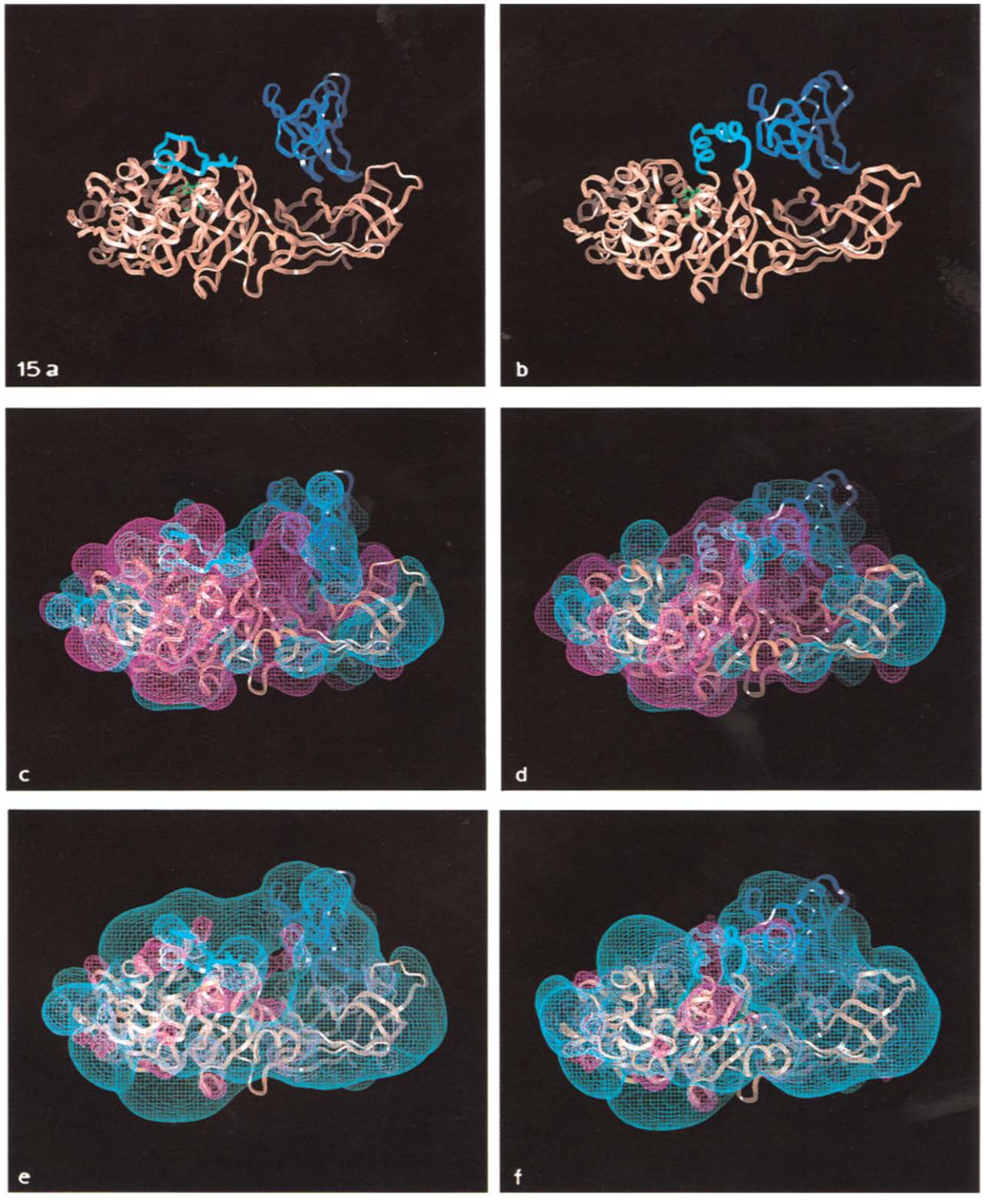
As an illustration of such type of comparisons, using pH-dependent electrostatics, we have applied the solvent accessibility-modified Tanford-Kirkwood method (see 6.2.) to the human pancreatic lipase structures with both closed (van Tilbeurgh et al., 1992) and open lid (van Tilbeurgh et al., 1993), as shown in Fig. 15a and b. Fig. 15c–f shows surfaces corresponding to an electrostatic potential equal to $\pm 1.0\ kT/e$ (where k is the Boltzman constant, T the absolute temperature and e the proton charge). These surfaces correspond to regions were the electrostatic interactions on a charge are roughly of the same magnitude as the thermal effects due to the surrounding solvent, i.e., where charged molecules in solution start to feel electrostatic steering or repulsion. At pH 7 clear differences exist between the closed and open forms, the latter showing a dipolar groove in the presumed binding site region. At pH 4 the molecule is strongly positively charged and most electrostatically differentiated regions have disappeared. Given the role of electrostatic interactions on molecular orientation and association (see the beginning of this section (6)), this is expected to markedly affect the interaction with the lipid-water interface.

For enzymes the catalytic activity involving a charged residue can be modulated by shifting the pK_a of that residue. The pK_a shifts of the active site histidine has been successfully predicted for a number of mutants of subtilisin (Gilson and Honig, 1987; Loewenthal et al., 1993).

One of the main reasons why enzymes are good catalysts is because they stabilise the transition state intermediate (Fersht, 1985). For enzymatic reactions that are not diffusion limited, engineering leading to an enhanced stabilisation of the intermediate will result in an increased activity. The induced dipole method was used to compute the activation free energy for different mutants of trypsin and subtilisin (Warshel et al., 1989), with some qualitative agreement with the experimental results.

The prediction of changes introduced by mutations on redox potentials could also be of interest to protein engineering. Prediction of redox potentials has been made with some success (Rogers et al., 1985; Durell et al., 1990). In plastocyanin the effect of chemically modifying charged groups was also considered (Durell et al., 1990). The effect of mutations could also be analysed, as has been done for pK_a shift calculations (see above).

The above examples clearly show that, whatever the particular method used, the modelling of


15 a
b
c
d
e
f

electrostatic interactions in proteins has an important role to play in protein engineering. A highly relevant example is the design of a faster 'perfect' enzyme (Getzoff et al., 1992), which also illustrates the combination of different methods (BD and electrostatic continuum methods) that can sometimes be determinant in a modelling study.

7. Protein engineering – future perspectives

The science of protein engineering is advancing rapidly, and is emerging in many new contexts, such as metabolic engineering. Rational protein engineering is a complex undertaking – and only the groups with sufficient understanding of sequences and 3-D structures can handle the complex underlying problems. Predicting protein structure may be difficult – but predicting future developments in a very active branch of science can be hazardous at the best. However, we will review a few of the more recent research aspects that we are convinced will be of key importance in the future development of protein engineering.

7.1. Non-conventional media

Often the substrates or products in an enzymatic process are poorly soluble in an aqueous medium. This may lead to poor yields and difficult or expensive purification steps. The potential of using other solvents, either pure or in mixture, where substrates and/or products may be soluble has attracted a great deal of attention (Tramper et al., 1992; Arnold, 1993).

Dissolving the protein in organic solvents will alter the macroscopic dielectric constant and lead to a much less pronounced difference between the interior and exterior static dielectric behaviour. Protein function in such media may be altered and is poorly understood; we can expect a significant development in the future.

Despite the often dramatic change in dielectric constant when changing the solvent from, e.g., water to an organic substance, the protein 3-D structure can remain virtually intact, as has been documented in the case of subtilisin Carlsberg dissolved in anhydrous acetonitrile (Fitzpatrick et al., 1993). The hydrogen bonding pattern of the active site environment is unchanged, and 99 of the 119 enzyme-bound structural water molecules are still in place. One-third of the 12 enzyme-bound acetonitrile molecules reside in the active site. Many enzymes remain active in organic solvents and in the case of enzyme reactions where the substrate has very poor water solubility, a change to organic solvent can be of major importance (Gupta, 1992).

An extreme case of a non-conventional medium for enzymatic action is the gas phase. Certain enzymes, immobilised on a solid bed, have been shown to be active at elevated temperatures towards selected substrates in the gas phase (Lamare and Legoy, 1993). Obviously the range of substrates that potentially can be used is limited to those that actually can be brought into the gas phase under conditions where the enzyme is still active. Enzymes for which such reactions have been studied include hydrogenase, alcohol oxidase and lipases. The fact that even interfacially activated lipases (such as the porcine pancreatic and the *Candida rugosa* lipases) function with gas phase carried substrate molecules opens up the interesting possibility of studying the role of water in this reaction.

Protein engineering may be used to enhance enzyme activity in organic solvents (Arnold, 1993;

Fig. 15. Electrostatic maps of HPL with closed and open lid. Ribbon models of human pancreatic lipase with colipase are shown with closed (left: a,c,e) and open (right: b,d,f) lid. The colipase is shown in blue and the mainly α-helical 'lid' region is highlighted in cyan. The residues of the active site are shown in green. Access to the active site pocket seems to be controlled by the conformational state of the lid. Electrostatic isopotential contours of $\pm 1.0\ kT/e$ are shown at pH 4 (c,d) and pH 7 (e,f). The negative surfaces are represented in red and the positive surfaces in blue. The models and isopotential contours were produced with *Insight II* and *DelPhi* (Biosym Technologies, San Diego). The pH-dependent charge sets were computed with TITRA (to be published).

Chen and Arnold, 1993). When dissolving subtilisin E in 60% dimethylformamide (DMF) the k_{cat}/K_M for the model substrate suc-Ala-Ala-Pro-Met-*p*-nitroanilide drops 333-fold. After ten mutations were introduced, the activity in DMF was restored almost to the level of the native enzyme in water.

7.2. Metabolic engineering

All metabolic conversions in micro-organisms are carried out directly or indirectly by proteins. Our ability to manipulate single genes has opened up for the actual *control* of such processes. We may alter the efficacy of a certain pathway or we may introduce totally new pathways. Thus, *Escherichia coli* can be modified in such a way that one can use D-glucose in the *E. coli* based manufacture of hydroquinone, benzoquinone, catechol and adipic acid (Dell and Frost, 1993; Draths and Frost, 1990; Frost, 1993). Presently such compounds are produced through organic chemical synthesis using aromatics as one of the reactants. The prospect of producing the same compounds using only microbes and glucose thus has some obvious environmental benefits. We expect to see a virtual surge in the engineering of micro-organisms towards the production of rare chemical or biochemical compounds or compounds for which the current synthetic route is costly either economically or from an environmental perspective.

7.3. De novo design

The perspective of designing and producing functional protein molecules from scratch is extremely attractive to many visionary scientists. Some central questions arise: Do we know enough to undertake such tasks, and what goals can we define? Screening mutation studies of protein interfaces show that the majority of mutations reduce activity or binding affinity (Cunningham and Wells, 1993), indicating that most proteins already represent highly optimised designs. The groups active in this area have aimed at constructing certain 3-dimensional folds such as the four helix bundle (Felix) (Hecht et al., 1990) and histidine-based metal binding sites (Arnold, 1993) and even the observation of limited enzymatic activity is regarded as a successful result (Johnson et al., 1993).

Protein de novo design of helix bundles may even follow a very simple binary pattern of polar and nonpolar amino acids as was concluded in a study of four-helix bundle proteins (Kamtekar et al., 1993). The helix–helix contact surfaces are mainly hydrophobic, whereas the solvent exposed regions are hydrophilic. Many variants conforming to this hydrophobic pattern were generated and two of these proteins were stabilised with 3.7 and 4.4 kcal mol^{-1} relatively to the unfolded form, thus approaching what is found for many natural proteins. The authors suggest that such a binary pattern may have been important in the early stages of evolution. In our laboratory we have results supporting this conclusion for the trypsin family of proteins, which is predominantly in a β-strand based fold (Petersen et al., 1994).

7.4. Hybrid proteins

Fusion and hybrid proteins may be produced by fusing the genes or gene fragments including a proper linking region between the two genes (Argos, 1990). This in principle may allow for combining properties from two different proteins. Thus artificial bifunctional enzymes have been produced by fusing the genes for the proteins, e.g., β-galactosidase and galactokinase (Bulow, 1990). In a recent paper an elegant hybrid protein concept is described. A hybrid antibody fragment was designed to consist of a heavy-chain variable domain from one antibody connected through a linker region of 5–15 residues to a short light-chain variable domain from another antibody (Holliger et al., 1993). The antibody fragments displayed similar binding characteristics as the parent antibodies. The prospect of engineering multifunctional antibodies for medical applications is imminent.

A hybrid protein between the glucose transporter and the *N*-acetylglucosamine transporter of *E. coli* have been produced. The two proteins displayed 40% residue identity. The hybrid protein consisted of the putative transmembrane do-

main from the glucose transporter and the two hydrophilic domains from the *N*-acetylglucosamine transporter. The hybrid protein was, somewhat surprisingly, still specific for glucose (Hummel et al., 1992). Interestingly, several naturally occurring proteins themselves seem to have originated through gene fusion. In the case of hexokinase it is proposed that it originated from a duplication of the glucokinase gene maintaining even the gene organisation (Kogure et al., 1993). Several other proteins such as receptor proteins of the insulin family can best be understood as gene fusion products of a kinase domain onto the rest of the receptor (which in itself may consist of several fragments).

With potential medical applications, protein-nucleic acid hybrids have been constructed, where the nucleic acid fragment complemented the sequence of a fragment of mRNA that the RNase should be targeted towards. The results obtained confirmed that this approach indeed worked (Kanaya et al., 1992). The potentials for generating anti-viral agents against, e.g., HIV are obvious.

7.5. Nano technology

As a consequence of the enormous growth in our understanding of molecular biology and material technology, a new technological sector is emerging which takes aim at exploring the possible advantages in creating micro-machines and switchable molecular entities. This concept is currently known as nano technology (Birge, 1992). Two concepts that we find particularly interesting are described briefly below.

7.5.1. Optical and chemical switches of molecular dimension

Rhodopsin is a very ancient molecular construct – we find rhodopsin like molecules in a range of roles, all of them associated with its membrane location. Proton transport and receptor functions are particularly interesting. Bacteriorhodopsin from *Halobacterium halobium* maintains a large pH gradient across the bacterial membrane. This protein complex is coloured, and its colour can be changed by exposing the protein to light of an appropriate frequency. The lifetime of the excited state can be adjusted by adjusting the physical chemical parameters of the medium the rhodopsin is embedded in (Birge, 1992). This protein can be used as a molecular switch in a very broad sense, e.g., as part of a high density memory device.

However, changing the colour of a protein molecule is just one example that could be considered. Another molecular based switch concept involves the transfer of a molecular ring (Paraquat-derived rotaxane ring) between two binding sites (Bradley, 1993). Currently the transfer is induced by a solvent change, but it is believed that an electrochemical transfer mechanism can be developed as well. Similar concepts can probably also be developed for proteins.

8. Conclusions

The present paper reviews some of many new developments in protein engineering. The review is not exhaustive – it is simply not possible to do this properly within the limits of this paper.

We have tried to review some selected scientific areas of key importance for protein engineering, such as the validity of protein sequence information as well as structural information. Sometimes the translation of a gene sequence to amino acid sequence is not trivial – a range of posttranscriptional editing and splicing events may occur, leading to a functional protein, where the amino acid sequence cannot be directly deducted from the gene sequence. In addition, posttranslational modification may provide triggers for other parts of the cells molecular machinery. We are thus in a situation where the full benefits and profits from projects such as the human genome project may escape us for a while.

We have covered some of the recent developments in the modelling of protein structure by homology, which we regard as one of the most strategic areas of development. We will be flooded with sequence information deducted from gene sequences, and in the cases where the deducted amino acid sequences are assumed valid, we have

to use homology based structure prediction in most cases. Given that the number of protein structure families is expected to be limited the task is durable. Here we should again caution the reader. We have no a priori reason to assume that non-soluble proteins, such as structural proteins, have structures that can be predicted from our limited library of mostly globular, soluble proteins. Some structural proteins are gigantic, the cuticle collagen in the *Riftia* worms from deep sea hydrothermal vents have a molecular mass of 2.600 kDa (Gaill et al., 1991). It is extremely unlikely that a 3-D structure at atomic resolution of such a protein will ever be determined using methods we have available today.

NMR has emerged with surprising speed as a structure determination tool. Many excellent reviews have been written on this topic. We have decided to direct the readers attention to some recent developments that we believe will be of significant importance to the usage of NMR in protein engineering projects. The potential of using NMR to study the solvent exposed outer shell of larger proteins, that by far exceed the 30 kDa limit mentioned earlier is intriguing. This is particularly so, since most functionality of a protein is a feature of exactly the residues in the outer shell. Thus, we can 'peel' the protein, and thereby isolate the spectral information that pertains to the surface only. This simplifies the spectra, and in some cases even allows for a partial assignment of specific residues.

Recent developments in pH-dependent protein electrostatics have been given special attention here. The similarities and differences within a family of structurally related proteins can only be understood if we are capable of interpreting the consequences of the substitutions, insertions and deletions that mostly occur at the surface of the proteins. When such changes are found and they involve charged residues, this will effect the extent or polarity of the electrostatic fields that the protein molecule is embedded in. We believe that the consequences of charge mutations to a large extent can be predicted through the use of pH-dependent electrostatics although practical examples are still lacking. To our knowledge the results on the electrostatic consequences of the lid motion in the human pancreatic lipase (vide supra) are among the first such reported.

The story of molecular biology is continuously unfolding – and our understanding of our own biology, development and evolution is becoming ever deeper and more detailed. But we are also, once again, discovering that one of the many qualities of Nature is endless complexity.

Acknowledgements

We want to thank Christian Cambillau, CNRS, Marseille, for kindly providing us with pre-release 3-D data of human pancreatic lipase, Jerry H. Brown, Harvard University, for sending us a pre-release dataset for the HLA II structure, Alwyn Jones, Uppsala University, for pre-release 3-D data of *Candida antarctica* B lipase, and John McCarthy, Brookhaven National Laboratory, for helping us with data on previous PDB releases. The French Norwegian Foundation (FNS 27958) and the Norwegian Research Council (BP 29345) have contributed with financial support to some of the research activities described in this paper. A.B. and P.M. thank Junta Nacional de Investigação Científica, Portugal, for their grants.

References

Abola, E., Bernstein, F.C., Bryant, S.H., Koetzle, T.F. and Weng, J. (1987). Protein Data Bank. Crystallographic databases – Information content, software systems, scientific applications. Bonn/Cambridge/Chester, Data Commission of the International Union of Crystallography. pp. 107–132.

Adler, B.K., Harris, M.E., Bertrand, K.I. and Hajduk, S.L. (1991) Modification of *Trypanosoma brucei* mitochondrial rRNA by posttranscriptional 3′ polyuridine tail formation. Mol. Cell Biol. 11(12), 5878–5884.

Alberts, B., Bray, D. Lewis, J., Raff, M., Roberts, K., Watson, J.D. (1983) Molecular Biology of the Cell. Garland Publ. Inc., New York.

Alexandrov, N. and Go, N. (1993) Significance of similarities in protein structures (in Abstracts of the 5th annual meeting of the Protein Engineering Society of Japan). Protein Eng. 6(8), 1003–1029.

Amrein, M. and Gross, H. (1992) Scanning tunneling microscopy of biological macromolecular structures coated with a conducting film. Scanning Microsc. 6(2), 335–343.

Argos, P. (1990) An investigation of oligopeptides linking domains in protein tertiary structures and possible candidates for general gene fusion. J. Mol. Biol. 211, 943–958.

Arnold, F.H. (1993) Engineering proteins for nonnatural environments. FASEB J. 7(9), 744–749.

Auger, M., McDermott, A.E., Robinson, V., Castelhano, A.L., Billedeau, R.J., Pliura, D.H., Krantz, A. and Griffin, R.G. (1993) Solid-state ^{13}C NMR study of a transglutaminase – inhibitor adduct. Biochemistry 32, 3930–3934.

Baca, M., Alewood, P.F. and Kent, S.B. (1993) Structural engineering of the HIV-1 protease molecule with a β-turn mimic of fixed geometry. Protein Sci. 2(7), 1085–1091.

Bairoch, A. and Boeckmann, B. (1992) The SWISS-PROT protein sequence data bank. Nucleic Acids Res. 20, 2019–2022.

Ball, P. (1994) Polymers made to measure. Nature 367, 323–324.

Barton, G.J. (1993) ALSCRIPT: A tool to format multiple sequence alignments. Protein Eng. 6(1), 37–40.

Bashford, D. and Karplus, M. (1990) pK_a's of ionizable groups in proteins: atomic detail from a continuum electrostatic model. Biochemistry 29, 10219–10225.

Bashford, D. and Karplus, M. (1991) Multiple-site titration curves of proteins: An analysis of exact and approximate methods for their calculation. J. Phys. Chem. 95, 9556–9561.

Bax, A. and Davis, D.G. (1985) MLEV-17-based two-dimensional homonuclear magnetization transfer spectroscopy. J. Magn. Reson. 65, 355–360.

Benner, S.A. and Gerloff, D.L. (1993) Predicting the conformation of proteins. Man versus machine. FEBS Lett. 325(1–2), 29–33.

Berg, O.G. and von Hippel, P.H. (1985) Diffusion-controlled macromolecular interactions. Annu. Rev. Biophys. Biophys. Chem. 14, 131–160.

Bernstein, F.C., Koetzle, T.F., Williams, G.J.B., Meyer, E.F., Brice, M.D., Rodgers, J.R., Kennard, O., Shimanouchi, T. and Tasumi, M. (1977) The Protein Data Bank: A computer-based archival file for macromolecular structures. J. Mol. Biol. 112, 535–542.

Beroza, P., Fredkin, D.R., Okamura, M.Y. and Feher, G. (1991) Protonation of interacting residues in a protein by a Monte Carlo method: Application to lysozyme and the photosynthetic reaction center of *Rhodobacter sphaeroides*. Proc. Natl. Acad. Sci. USA 88, 5804–5808.

Beveridge, D.L. and Dicapua, F.M. (1989) Free energy via molecular simulation: Application to chemical and biomolecular systems. Annu. Rev. Biophys. Biophys. Chem. 18, 431–492.

Birge, R. (1992) Molecular electronics. In: Crandall, B.C and Lewis, J. (Eds.) Nanotechnology. Research and Perspectives. MIT Press, Cambridge, MA.

Blundell, T.L. and Johnson, M.S. (1993) Catching a common fold. Protein Sci. 2(6), 877–883.

Böck, A., Forschhammer, K. Heider, J. and Baron, C. (1991) Seleno protein synthesis: An expansion of the genetic code. Trends Biochem. Sci. 16, 463–467.

Bohr, J., Bohr, H., Brunak, S., Cotterill, R.M.J., Fredholm, H., Lautrup, B. and Petersen, S.B. (1993) Protein structures from distance inequalities. J. Mol. Biol. 231, 861–869.

Boscott, P.E., Barton, G.J. and Richards, W.G. (1993) Secondary structure prediction for modelling by homology. Protein Eng. 6(3), 261–266.

Bowie, J.U. and Eisenberg, D. (1993) Inverted protein structure prediction. Curr. Opin. Struct. Biol. 3, 437–444.

Bowie, J.U., Lüthy, R. and Eisenberg, D. (1991) A method to identify protein sequences that fold into a known three-dimensional structure. Science 253, 164–170.

Bradley, D. (1993) Will future computers be all wet? Science 259, 890–892.

Braunschweiler, L. and Ernst, R.R. (1983) Coherence transfer by isotropic mixing: Application to proton correlation spectroscopy. J. Magn. Reson. 53, 521–528.

Broadhurst, R.W., Dobson, C.M., P.J.Hore, Radford, S.E. and Rees, M.L. (1991) A photochemically induced dynamic nuclear polarization study of denatured states of lysozyme. Biochemistry 30, 405–412.

Brown, J.H., Jardetzky, T.S., Gorga, J.C., Stern, L.J., Urban, R.G., Strominger, J.L. and Wiley, D.C. (1993) Three-dimensional structure of the human class II histocompatibility antigen HLA-DR1. Nature 364(6432), 33–39.

Bryant, S.H. and Lawrence, C.E. (1993) An empirical energy function for threading protein sequence through the folding motif. Proteins 16(1), 92–112.

Bulow, L. (1990) Preparation of artificial bifunctional enzymes by gene fusion. Biochem. Soc. Symp. 57, 123–133.

Burke, P.A., Griffin, R.G. and Klibanov, A.M. (1992) Solid-state NMR assessment of enzyme active center structure under nonaqueous conditions. J. Biol. Chem. 267, 20057–20064.

Burley, S.K. (1994) Forward to the fundamentals. Struct. Biol. 1(1), 8–10.

Cassels, R., Dobson, C.M., Poulsen, F.M. and Williams, R.J.P. (1978) Study of the tryptophan residues of lysozyme using ^1H nuclear magnetic resonance. Eur. J. Biochem. 95, 81–97.

Cattaneo, R. (1994) RNA duplexes guide base conversions. Curr. Biol. 4, 134–136.

Chang, C.A., Brittain, H.G., Telser, J. and Tweedle, M.F. (1990) pH dependence of relaxivities and hydration numbers of Gadolinium(III) complexes of linear amino carboxylates. Inorg. Chem. 29, 4468–4473.

Chazin, W.J., Hugli, T.E. and Wright, P.E. (1988) ^1H NMR studies of human C3a anaphylatoxin in solution: Sequential resonance assignments, secondary structure, and global fold. Biochemistry 27, 9139–9148.

Chen, K. and Arnold, F.H. (1993) Tuning the activity of an enzyme for unusual environments: Sequential random mutagenesis of subtilisin E for catalysis in dimethylformamide. Proc. Natl. Acad. Sci. USA 90(12), 5618–5622.

Chothia, C. (1992) Proteins. One thousand families for the molecular biologist. Nature 357(6379), 543–544.

Chou, K.C. and Zhang, C.T. (1992) A correlation-coefficient method to predicting protein-structural classes from amino acid compositions. Eur. J. Biochem. 207(2), 429–433.

Christensen, A.M. and Schaefer, J. (1993) Solid-state NMR

determination of intra- and intermolecular ^{31}P-^{13}C distances for shikimate 3-phosphate and [1-^{13}C]glyphosate bound to enolpyruvylshikimate-3-phosphate synthase. Biochemistry 32, 2868–2873.

Clore, G., Kay, L., Bax, A. and Gronenborn, A. (1991a) Four-dimensional $^{13}C/^{13}C$-edited nuclear Overhauser enhancement spectroscopy of a protein in solution: Application to interleukin 1β. Biochemistry 30, 12–18.

Clore, G.M., Wingfield, P.T. and Gronenborn, A.M. (1991b) High-resolution three-dimensional structure of interleukin 1β in solution by three- and four-dimensional nuclear magnetic resonance spectroscopy. Biochemistry 30(9), 2315–2323.

Cohen, B.I., Presnell, S.R. and Cohen, F.E. (1993) Origins of structural diversity within sequentially identical hexapeptides. Protein Sci. 2, 2134–2145.

Colloc'h, N., Etchebest, C., Thoreau, E., Henrissat, B. and Mornon, J.P. (1993) Comparison of three algorithms for the assignment of secondary structure in proteins: The advantages of a consensus assignment. Protein Eng. 6(4), 377–382.

Coulson, A. (1993) Extracting the information – Sequence analysis software design evolves. Trends Biotechnol. 11, 223–227.

Covell, D.G. and Jernigan, R.L. (1990) Conformations of folded proteins in restricted spaces. Biochemistry 29(13), 3287–3294.

Crippen, G.M. (1991) Prediction of protein folding from amino acid sequence over discrete conformation spaces. Biochemistry 30(17), 4232–4237.

Cunningham, B.C. and Wells, J.A. (1993) Comparison of a structural and a functional epitope. J. Mol. Biol. 233, 554–563.

Davies, M.E. and McCammon, J.A. (1990) Electrostatics in biomolecular structure and dynamics. Chem. Rev. 90, 509–521.

Dell, K.A. and Frost, J.W. (1994) Identification and removal of impediments to biocatalytic synthesis of aromatics from D-glucose: Rate-limiting enzymes in the common pathway of aromatic amino acid biosynthesis. J. Am. Chem. Soc., in press.

Derewenda, Z.S., Derewenda, U. and Dodson, G.G. (1992) The crystal and molecular structure of the *Rhizomucor miehei* triacylglyceride lipase at 1.9 Å resolution. J. Mol. Biol. 227(3), 818–839.

Diamond, R. (1974) Real-space refinement of the structure of hen egg white lysozyme. J. Mol. Biol. 82, 371.

Dill, K.A. (1990) Dominant forces in protein folding. Biochemistry, 29, 7133–7155.

Dobson, C.M., Ferguson, S.J., Poulsen, F.M. and Williams, R.J.P. (1978) Complete assignment of aromatic ^{1}H nuclear magnetic resonances of the tyrosine residues of hen lysozyme. Eur. J. Biochem. 92, 99–103.

Doolittle, R.F. (1992) Stein and Moore Award address. Reconstructing history with amino acid sequences. Protein Sci. 1(2), 191–200.

Doolittle, R.F. (1993) The comings and goings of homing endonucleases and mobile introns. Proc. Natl. Acad. Sci. USA 90(12), 5379–5381.

Drabløs, F. and Petersen, S.B. (1994) Multim – Tools for multiple sequence analysis. In preparation.

Draths, K.M. and Frost, J.W. (1990) Genomic direction of synthesis during plasmid-based biocatalysis. J. Am. Chem. Soc. 112, 9630–9632.

Drayney, D. and Kingsbury, C.A. (1981) Free radical induced nuclear magnetic resonance shifts: Comments on contact shift mechanism. J. Am. Chem. Soc. 103, 1041–1047.

Dubchak, I., Holbrook, S.R. and Kim, S.H. (1993) Prediction of protein folding class from amino acid composition. Proteins Struct. Func. Genet. 16(1), 79–91.

Durell, S.R., Labanowski, J.K. and Gross, E.L. (1990) Modeling of the electrostatic potential field of plastocyanin. Arch. Biochem. Biophys. 277, 241–254.

Eisenberg, D., Bowie, J.U., Luthy, R. and Choe, S. (1992) Three-dimensional profiles for analysing protein sequence – structure relationships. Faraday Discuss. 1992(93), 25–34.

Eisenmenger, F., Argos, P. and Abagyan, R. (1993) A method to configure protein side-chains from the main-chain trace in homology modelling. J. Mol. Biol. 231(3), 849–860.

Emsley, J., White, H.E., O'Hara, B.P., Oliva, G., Srinivasan, N., Tickle, I.J., Blundell, T.L., Pepys, M.B. and Wood, S.P. (1994) Structure of pentameric human serum amyloid P component. Nature 367, 338–345.

Ernst, R.R. (1992) Nuclear magnetic resonance fourier transform spectroscopy (Nobel lecture). Angew. Chem. 31, 805–930.

Esposito, G., Lesk, A.M., Molinari, H., Motta, A., Niccolai, N. and Pastore, A. (1992) Probing protein structure by solvent pertubation of nuclear magnetic resonance spectra. J. Mol. Biol. 224, 659–670.

Fahy, G.M. (1993) Molecular nanotechnology. Clin. Chem. 39(9), 2011–2016.

Fairbrother, W.J., Gippert, G.P., Reizer, J., Saier, M.J. and Wright, P.E. (1992) Low resolution solution structure of the *Bacillus subtilis* glucose permease IIA domain derived from heteronuclear three-dimensional NMR spectroscopy. FEBS Lett. 296(2), 148–152.

Farabaugh, P.J. (1993) Alternative readings of the genetic code. Cell 74(4), 591–596.

Fersht, A. (1985) Enzyme Structure and Mechanism. Freeman, New York.

Fersht, A. and Winter, G. (1992) Protein engineering. Trends Biochem. Sci. 17(8), 292–295.

Fitzpatrick, P.A., Steinmetz, A.C., Ringe, D. Klibanov, A.M. (1993) Enzyme crystal structure in a neat organic solvent. Proc. Natl. Acad. Sci. USA 90, 8653–8657.

Foght, R.H., Schipper, D., Boelens, R. and Kaptein, R. (1994) ^{1}H, ^{13}C and ^{15}N NMR backbone assignments of the 269-residue serine protease PB92 from *Bacillus alcalophilus*. J. Biomol. NMR 4, 123–128.

Frey, M.H., Wagner, G., Vasak, M., Sørensen, O.W., Neuhaus, D., Worgotter, E., Kagi, J.H.R., Ernst, R.R. and Wüthrich, K. (1985) Polypeptide – metal cluster connectivities in metallothionein 2 by novel ^{1}H-^{113}Cd heteronuclear two-di-

mensional NMR experiments. J. Am. Chem. Soc. 107, 6847–6851.

Frost, J.W. (1993) Design and use of heterologous microbes for conversion of D-glucose into aromatic chemicals. Enzyme engineering XII, Deauville, France, September 19–24, 1993.

Gaill, F., Wiedemann, H., Mann, K., Kuhn, K., Timpl, R. and Engel, J. (1991) Molecular characterization of the cuticle and interstitial collagens from worms collected at deep sea hydrothermal vents. J. Mol. Biol. 221, 209–223.

George, D.G., Barker, W.C. and Hunt, L.T. (1986) The protein identification resource (PIR). Nucleic Acids Res. 14(1), 11–15.

Getzoff, E.D., Cabelli, D.E., Fisher, C.L., Parge, H.E., Viezzoli, M.S., Banci, L. and Hallewell, R.A. (1992) Faster superoxide dismutase mutants designed by enhancing electrostatic guidance. Nature 358(6384), 347–351.

Ghadiri, M.R., Granja, J.R., Milligan, R.A., McRee, D.E. and Khazanovich, N. (1993) Self-assembling organic nanotubes based on a cyclic peptide architecture. Nature 366, 324–327.

Gilson, M.K. (1993) Multiple-site titration and molecular modeling: Two rapid methods for computing energies and forces for ionizable groups in proteins. Proteins 15, 266–282.

Gilson, M.K. and Honig, B. (1988) Calculation of the total electrostatic energy of a macromolecular system: Solvation energies, binding energies, and conformational analysis. Proteins Struct. Funct. Genet. 4, 7–18.

Gilson, M.K. and Honig, B. (1991) The inclusion of electrostatic hydration energies in molecular mechanics calculations. J. Computer-Aided Mol. Design 5, 5–20.

Gilson, M.K. and Honig, B.H. (1987) Calculations of electrostatic potentials in an enzyme active site. Nature 330, 84–86.

Gilson, M.K., Sharp, K.A. and Honig, B.H. (1987) Calculating the electrostatic potential of molecules in solution: Method and error assessment. J. Comp. Chem. 9, 327–335.

Gracy, J., Chiche, L. and Sallantin, J. (1993) Improved alignment of weakly homologous protein sequences using structural information. Protein Eng. 6(8), 821–829.

Gray, M.W. and Covello, P.S. (1993) RNA editing in plant mitochondria and chloroplasts. FASEB J. 7(1), 64–71.

Green, H., (1993) Human genetic diseases due to codon reiteration: Relationship to an evolutionary mechanism. Cell 74, 955–956.

Gregory, R.B., Gangoda, M., Gilpin, R.K. and Su, W. (1993) The influence of hydration on the conformation of lysozyme studied by solid-state ^{13}C-NMR spectroscopy. Biopolymers 33, 513–519.

Griesinger, C., Sørensen, O.W. and Ernst, R.R. (1989) Three-dimensional fourier spectroscopy. Application to high-resolution NMR.J. Magn. Reson. 84, 14–63.

Grivell, L.A. (1994) Invasive introns. Curr. Biol. 4, 161–164.

Gupta, M.N. (1992) Enzyme function in organic solvents. Eur. J. Biochem. 203(1–2), 25–32.

Haggerty, L. and Lenhoff, A.M. (1993) Analysis of ordered arrays of adsorbed lysozyme by scanning tunneling microscopy. Biophys. J. 64(3), 886–895.

Harris, M., Decker, C., Sollner, W.B. and Hajduk, S. (1992) Specific cleavage of pre-edited mRNAs in trypanosome mitochondrial extracts. Mol. Cell Biol. 12(6), 2591–2598.

Harvey, S.C. (1989) Treatment of electrostatic effects in macromolecular modeling. Proteins Struct. Func. Genet. 5, 78–92.

Hecht, M.H., Richardson, J.S., Richardson, D.C. and Odgen, R.C. (1990). De novo design, expression and characterization of Felix: A four-helix bundle protein of native like sequence. Science 249, 884–891.

Hedstrom, L., Szilagyi, L. and Rutter, W.J. (1992) Converting trypsin to chymotrypsin: The role of surface loops. Science 255(5049), 1249–1253.

Hendlich, M., Lackner, P., Weitckus, S., Floeckner, H., Froschauer, R., Gottsbacher, K., Casari, G. and Sippl, M.J. (1990) Identification of native protein folds amongst a large number of incorrect models. The calculation of low energy conformations from potentials of mean force. J. Mol. Biol. 216(1), 167–180.

Hernandez, G., Brittain, H.G., Tweddle, M.F. and Bryant, R.G. (1990a) Nuclear magnetic relaxation in aqueous solutions of the Gd(HEDTA) complex. Inorg. Chem. 29, 985–988.

Hernandez, G., Tweedle, M.F. and Bryant, R.G. (1990b) Proton magnetic relaxation dispersion in aqueous glycerol solutions of $Gd(DTPA)^{2-}$ and $Gd(DOTA)^{-}$. Inorg. Chem. 29, 5109–5113.

Higaki, J.N., Fletterick, R.J. and Craik, C.S. (1992) Engineered metalloregulation in enzymes. Trends Biochem. Sci. 17(3), 100–104.

Higuchi, M., Single, F.N., Köhler, M., Sommer, B., Sprengel, R. and Seeburg, P.H. (1993) RNA editing of AMPA receptor subunit GluR-B: A base-paired intron-exon structure determines position and efficiency. Cell 75, 1361–1370.

Hodges, R.A., Perler, F.B., Noren, C.J. and Jack, W.E. (1992) Protein splicing removes intervening sequences in an archaea DNA polymerase. Nucleic Acids Res. 20(23), 6153–6157.

Hol, W.G.J. (1985) The role of the α-helix dipole in protein function and structure. Prog. Biophys. Mol. Biol. 45, 149–195.

Holliger, P., Prospero, T. and Winter, G. (1993) 'Diabodies': small bivalent and bispecific antibody fragments. Proc. Natl. Acad. Sci. USA 90, 6444–6448.

Holm, L. and Sander, C. (1993) Globin fold in a bacterial toxin. Nature 361(6410), 309.

Holm, L., Ouzounis, C., Sander, C., Tuparev, G. and Vriend, G. (1992) A database of protein structure families with common folding motifs. Protein Sci. 1(12), 1691–1698.

Hore, P.J. and Kaptein, R. (1983) Proton nuclear magnetic resonance assignment and surface accessibility of tryptophan residues in lysozyme using photochemically induced dynamic nuclear polarization spectroscopy. Biochemistry 22, 1906–1911.

Hummel, U. Nuoffer, C., Zanolari, B. and Erni, B. (1992). A functional protein hybrid between the glucose transporter and the *N*-acetylglucosamine transporter of *Escherichia coli*. Protein Sci. 1, 356–362.

Jackson, J.D. (1975) Classical electrodynamics. John Wiley & Sons, New York.

Johnsson, K., Allemann, R.K., Widmer, H. and Benner, S.A. (1993) Synthesis, structure and activity of artificial, rationally designed catalytic polypeptides. Nature 365(6446), 530–532.

Jones, D.T., Taylor, W.R. and Thornton, J.M. (1992) A new approach to protein fold recognition. Nature 358(6381), 86–89.

Kaarsholm, N.C., Norris, K., Jorgensen, R.J., Mikkelsen, J., Ludvigsen, S., Olsen, O.H., Sorensen, A.R. and Havelund, S. (1993) Engineering stability of the insulin monomer fold with application to structure-activity relationships. Biochemistry 32(40), 10773–10778.

Kabsch, W. and Sander, C. (1983) Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 22, 2577–2637.

Kamtekar, S., Schiffer, J.M., Xiong, H., Babik, J.M. and Hecht, M.H. (1993) Protein design by binary patterning of polar and nonpolar amino acids. Science 262, 1680–1685.

Kanaya, S., Nakai, C., Konishi, A., Inoue, H., Ohtsuka, E. and Ikehara, M. (1992) A hybrid ribonuclease H. A novel RNA cleaving enzyme with sequence-specific recognition. J. Biol. Chem. 267, 8492–8498.

Kay, E.L., Clore, G.M., Bax, A. and Gronenborg, A.M. (1990) Four-dimensional heteronuclear triple-resonance NMR spectroscopy of interleukin-1β in solution. Science 249, 411–414.

Kessler, H., Gehrke, M. and Griesinger, C. (1988) Two-dimensional spectroscopy: Background and overview of the experiments. Angew. Chem. Int. Ed. Engl. 27, 490–536.

Killian, J.A., Taylor, M.J. and Koeppe, R.E. (1992) Orientation of the valine-1 side chain of the gramicidin transmembrane channel and implications for channel functioning. A ^{2}H NMR study. Biochemistry 31, 11283–11290.

Kim, J.L., Nikolov, D.B. and Burley, S.K. (1993a) Co-crystal structure of TBP recognizing the minor groove of a TATA element. Nature 365, 520–527.

Kim, Y., Geiger, J.H., Hahn, S. and Sigler, P.B. (1993b) Crystal structure of a yeast TBP/TATA-box complex. Nature 365, 512–520.

Klevit, R.E. and Waygood, E.B. (1986) Two-dimensional ^{1}H NMR studies of histidine-containing protein from *Escherichia coli*. Secondary and tertiary structure as determined by NMR. Biochemistry 25, 7774–7781.

Klimasauskas, S., Kumar, S., Roberts, R.J. and Cheng, X. (1994) *Hha*I methyltransferase flips its target base out of the DNA helix. Cell 76, 357–369.

Knegtel, R.M., Katahira, M., Schilthuis, J.G., Bonvin, A.M., Boelens, R., Eib, D., van der Saag, P.T. and Kaptein, R. (1993) The solution structure of the human retinoic acid receptor-β DNA-binding domain. J. Biomol. NMR 3(1), 1–17.

Kobe, B. and Deisenhofer, J. (1993) Crystal structure of porcine ribonuclease inhibitor, a protein with leucine-rich repeats. Nature 366, 751–756.

Kogure, K., Shinohara, Y. and Terada, H. (1993) Evolution of the type II hexokinase gene by duplication and fusion of the glucokinase gene with conservation of its organization. J. Biol. Chem. 268(12), 8422–8424.

Köhler, M., Burnashev, N., Sakmann, B. and Seeburg, P.H. (1993) Determinants of Ca^{2+} permeability in both TM1 and TM2 of high affinity kainate receptor channels: diversity by RNA editing. Neuron 10(3), 491–500.

Kohlstaedt, L.A., Wang, J., Friedman, J.M., Rice, P.A. and Steitz, T.A. (1992) Crystal structure at 3.5 Å resolution of HIV-1 reverse transcriptase complexed with an inhibitor. Science 256(5065), 1783–1790.

Koppenol, W.H. and Margoliash, E. (1982) The asymmetric distribution of charges on the surface of horse cytochrome c. J. Biol. Chem. 257, 4426–4437.

Kraulis, P.J. (1991) Molscript: A program to produce both detailed and schematic plots of protein structures. J. Appl. Cryst. 24, 946–950.

Kühlbrandt, W., Wang, D.N. and Fujiyoshi, Y. (1994) Atomic model of plant light-harvesting complex by electron crystallography. Nature 367, 614–621.

Lamare, S. and Legoy, M.-D. (1993) Biocatalysis in the gas phase. Trends Biotechnol. 11, 413–418.

Laskowski, R.A., MacArthur, M.W., Moss, D.S. and Thornton, J.M. (1993) PROCHECK: A program to check the stereochemical quality of protein structures. J. Appl. Cryst. 26, 283–291.

Lessel, U. and Schomburg, D. (1993) A new procedure for the detection and evaluation of similar substructures in proteins (in Abstracts of the 5th annual meeting of the Protein Engineering Society of Japan). Protein Eng. 6(8), 1003–1029.

Levin, J.M., Pascarella, S., Argos, P. and Garnier, J. (1993) Quantification of secondary structure prediction improvement using multiple alignments. Protein Eng. 6(8), 849–854.

Levitt, M. (1989) Molecular dynamics of macromolecules in water. Chemica Scripta 29A, 197–203.

Lewerenz, H.J., Jungblut, H., Campbell, S.A., Giersig, M. and Müller, D.J. (1992) Direct observation of reverse transcriptases by scanning tunneling microscopy. Aids Res. Hum. Retroviruses 8(9), 1663–1667.

Linderstrøm-Lang, K. (1924) On the ionization of proteins. C.R. Trav. Lab. Carlsberg 15, 1–29.

Loewenthal, R., Sancho, J., Reinikainen, T. and Fersht, A. (1993) Long-range surface charge-charge interactions in proteins. J. Mol. Biol. 232, 574–583.

Lüthy, R., Bowie, J.U. and Eisenberg, D. (1992) Assessment of protein models with three-dimensional profiles. Nature 356, 83–85.

Lüthy, R., Xenarios, I. and Bucher, P. (1994) Improving the

sensitivity of the sequence profile method. Protein Sci. 3, 139–146.

Luty, B.A., Wade, R.C., Madura, J.D., Davis, M.E., Briggs, J.M. and McCammon, J.A. (1993) Brownian dynamics simulations of diffusional encounters between triosephosphate isomerase and glyceraldehyde phosphate: electrostatic steering of glyceraldehyde phosphate. J. Phys. Chem. 97, 233–237.

Mark, A.E., Berendsen, H.J.C. and van Gunsteren, W.F. (1991) Conformational flexibility of aqueous monomeric and dimeric insulin: a molecular dynamics study. Biochemistry 30, 10866–10872.

Martin, J.L., Bardwell, J.C. and Kuriyan, J. (1993) Crystal structure of the DsbA protein required for disulphide bond formation in vivo. Nature 365(6445), 464–468.

Matthew, J.B. (1985) Electrostatic effects in proteins. Annu. Rev. Biophys. Biophys. Chem. 14, 387–417.

McCammon, J.A. and Harvey, S.C. (1987) Dynamics of Proteins and Nucleic Acids. Cambridge University Press, Cambridge, UK.

McDowell, L.M., Holl, S.M., Qian, S.J., Li, E. and Schaefer, J. (1993) Inter-tryptophan distances in rat cellular retinol binding protein II by solid-state NMR. Biochemistry 32, 4560–4563.

McKie, J.H., Jaouhari, R., Douglas, K.T., Goffner, D., Feuillet, C., Grima, P.J., Boudet, A.M., Baltas, M. and Gorrichon, L. (1993) A molecular model for cinnamyl alcohol dehydrogenase, a plant aromatic alcohol dehydrogenase involved in lignification. Biochim. Biophys. Acta 1202(1), 61–69.

Moxon, E.R., Rainey, P.B., Nowak, M.A. and Lenski, R.E. (1994) Adaptive evolution of highly mutable loci in pathogenic bacteria. Curr. Biol. 4(1), 24–33.

Namboodiri, K., Pattabiraman, N., Lowrey, A. and Gaber, B.P. (1988) Automated protein structure data bank similarity searches and their use in molecular modeling with MIDAS.J. Mol. Graphics 6, 211–212.

Nishikawa, K. and Matsuo, Y. (1993) Development of pseudoenergy potentials for assessing protein 3-D–1-D compatability and detecting weak homologies. Protein Eng. 6(8), 811–820.

Northrup, S.H., Boles, J.O. and Reynolds, J.C.L. (1988) Brownian dynamics of cytochrome *c* and cytochrome *c* peroxidase electron transfer proteins. Science 241, 67–70.

Northrup, S.H., Pear, M.R., Morgan, J.D., McCammon, J.A. and Karplus, M. (1981) Molecular dynamics of ferrocytochrome *c*. Magnitude and anisotropy of atomic displacements. J. Mol. Biol. 153, 1087–1109.

Novotny, J., Bruccoleri, R. and Karplus, M. (1984) An analysis of incorrectly folded protein models. Implications for structure predictions. J. Mol. Biol. 177(4), 787–818.

Omer, C.A., Kral, A.M., Diehl, R.E., Prendergast, G.C., Powers, S. Allen, C.M., Gibbs, J.B. and Kohl, N.E. (1993) Characterization of recombinant human farnesyl-protein transferase: Cloning, expression, farnesyl diphosphate binding, and functional homology with yeast prenyl-protein transferases. Biochemistry 32, 5167–5176.

Orengo, C.A., Brown, N.P. and Taylor, W.R. (1992) Fast structure alignment for protein databank searching. Proteins 14(2), 139–167.

Orengo, C.A., Flores, T.P., Taylor, W.R. and Thornton, J.M. (1993) Identification and classification of protein fold families. Protein Eng. 6(5), 485–500.

Orttung, W.H. (1977) Direct solution of the Poisson equation for biomolecules of arbitrary shape, polarizability density, and charge distribution. Ann. NY Acad. Sci. 303, 22–37.

Ouzounis, C., Sander, C., Scharf, M. and Schneider, R. (1993) Prediction of protein structure by evaluation of sequence-structure fitness. Aligning sequences to contact profiles derived from three-dimensional structures. J. Mol. Biol. 232(3), 805–825.

Overington, J., Donnelly, D., Johnson, M.S., Šali, A. and Blundell, T.L. (1992) Environment-specific amino acid substitution tables: Tertiary templates and prediction of protein folds. Protein Sci. 1(2), 216–226.

Pawson, T. (1992) SH2 and SH3 domains. Curr. Opin. Struct. Biol. 2, 432–437.

Pearson, W.R. (1990) Rapid and sensitive sequence comparison with FASTP and FASTA. Methods Enzymol. 183(63), 63–98.

Pearson, W.R. and Lipman, D.J. (1988) Improved tools for biological sequence comparison. Proc. Natl. Acad. Sci. USA 85(8), 2444–2448.

Petersen, S.B., Hough, E. and Baptista, A. (1994) Gene duplication and the origin of trypsin. In preparation.

Petersen, S.B. and Martel, P. (1994) Protein Engineering – New or Improved Proteins for Mankind. In: Cabral, J., Best, D., Boross, L. and Tramper, J. (Eds.) Applied Biocatalysis. Harwood Academic Publishers. In press.

Petros, A.M., Mueller, L. and Kopple, K.D. (1990) NMR identification of protein surfaces using paramagnetic probes. Biochemistry 29, 10041–10048.

Pitts, J.E., Dhanaraj, V., Dealwis, C.G., Mantafounis, D., Nugent, P., Orprayoon, P., Cooper, J.B., Newman, M. and Blundell, T.L. (1992) Multidisciplinary cycles for protein engineering: Site-directed mutagenesis and X-ray structural studies of aspartic proteinases. Scand. J. Clin. Lab. Invest. Suppl. 210(39), 39–50.

Rao, S., Zhu, Q.L., Vajda, S. and Smith, T. (1993) The local information content of the protein structural database. FEBS Lett. 322(2), 143–146.

Rayment, I., Holden, H.M., Whittaker, M., Yohn, C.B., Lorenz, M., Holmes, K.C. and Milligan, R.A. (1993) Structure of the actin – myosin complex and its implications for muscle contraction. Science 261(5117), 58–65.

Read, L.K., Myler, P.J. and Stuart, K. (1992) Extensive editing of both processed and preprocessed maxicircle CR6 transcripts in *Trypanosoma brucei*. J. Biol. Chem. 267(2), 1123–1128.

Redfield, C. and Dobson, C.M. (1988) Sequential ^{1}H-NMR

assignments and secondary structure of hen egg white lysozyme in solution. Biochemistry 27, 122–136.

Reynolds, J.C.L., Cooke, K.F. and Northrup, S.H. (1990) Electrostatics and diffusional dynamics in the carbonic anhydrase active site channel. J. Phys. Chem. 94, 985–991.

Richards, F.M. and Kundrot, C.E. (1988) Identification of structural motifs from protein coordinate data: Secondary structure and first-level supersecondary structure. Proteins 3(2), 71–84.

Ring, C.S. and Cohen, F.E. (1993) Modeling protein structures: Construction and their applications. FASEB J. 7(9), 783–790.

Roberts, G.C.K. (1993) NMR of macromolecules. A practical approach. In: D. Rickwood and B.D. Hames (Eds.) The Practical Approach Series. Vol. 134, Oxford University Press Inc., New York.

Rogers, N.K. (1986) The modelling of electrostatic interactions in the function of globular proteins. Prog. Biophys. Mol. Biol. 48, 37–66.

Rogers, N.K., Moore, G.R. and Sternberg, M.J.E (1985) Electrostatic interactions in globular proteins: Calculation of the pH dependence of the redox potential of cytochrome C_{551}. J. Mol. Biol. 182, 613–616.

Rooman, M.J. and Wodak, S.J. (1992) Extracting information on folding from the amino acid sequence: Consensus regions with preferred conformation in homologous proteins. Biochemistry 31(42), 10239–10249.

Rost, B. and Sander, C. (1993a) Prediction of protein secondary structure at better than 70% accuracy. J. Mol. Biol. 232(2), 584–599.

Rost, B. and Sander, C. (1993b) Secondary structure prediction of all-helical proteins in two states. Protein Eng. (8), 831–836.

Rost, B., Sander, C. and Schneider, R. (1994) PHD – An automatic mail server for protein secondary structure prediction. Comput. Appl. Biosci. 10(1), 53.

Rost, B., Schneider, R. and Sander, C. (1993) Progress in protein structure prediction? Trends Biochem. Sci. 18(4), 120–123.

Salzberg, S. and Cost, S. (1992) Predicting protein secondary structure with a nearest-neighbor algorithm. J. Mol. Biol. 227(2), 371–374.

Sander, C. and Schneider, R. (1991) Database of homology-derived protein structures and the structural meaning of sequence alignment. Proteins Struct. Funct. Genet. 9, 56–68.

Scheffler, J.E., Cottrell, C.E. and Berliner, L.J. (1985) An winexpensive, versatile sample illuminator for photo-CIDNP on any NMR spectrometer. J. Magn. Reson. 63, 199–201.

Schrag, J.D., Winkler, F.K. and Cygler, M. (1992) Pancreatic lipases: Evolutionary intermediates in a positional change of catalytic carboxylates? J. Biol. Chem. 267(7), 4300–4303.

Schuler, G.D., Altschul, S.F. and Lipman, D.J. (1991) A workbench for multiple alignment construction and analysis. Proteins 9(3), 180–190.

Shakhnovich, E.I. and Gutin, A.M. (1993) A new approach to the design of stable proteins. Protein Eng. 6(8), 793–800.

Sharp, K.A. and Honig, B. (1990) Electrostatic interactions in macromolecules: Theory and applications. Annu. Rev. Biophys. Biophys. Chem. 19, 301–332.

Sheridan, R.P. and Allen, L.C. (1980) The electrostatic potential of the alpha helix. Biophys. Chem. 11, 133–136.

Shire, S.J., Hanania, G.I.H. and Gurd, F.R.N (1974) Electrostatic effects in myoglobin. Hydrogen ion equilibria in sperm whale ferrimyoglobin. Biochemistry 13, 2967–2974.

Sines, J.J., Allison, S.A. and McCammon, J.A. (1990) Point charge distributions and electrostatic steering in enzyme/substrate encounter: Brownian Dynamics of modified copper/zinc superoxide dismutases. Biochemistry 29, 9403–9412.

Sippl, M.J. (1993a) Boltzmann's principle, knowledge based mean fields and protein folding. J. Computer-aided Mol. Design 7, 473–501.

Sippl, M.J. (1993b) Recognition of errors in three-dimensional structures of proteins. Proteins Struct. Funct. Genet. 17, 355–362.

Sklenar, H., Etchebest, C. and Lavery, R. (1989) Describing protein structure: A general algorithm yielding complete helicoidal parameters and a unique overall axis. Proteins Struct. Funct. Genet. 6(1), 46–60.

Solmajer, T. and Mehler, E.L. (1991) Electrostatic screening in molecular dynamics simulations. Protein Eng. 4, 911–917.

Soman, K., Yang, A.S., Honig, B. and Fletterick, R. (1989) Electrical potentials in trypsin isozymes. Biochemistry 28, 9918–9926.

Sommer, B., Köhler, M., Sprengel, R. and Seeburg, P.H. (1991) RNA editing in brain controls a determinant of ion flow in glutamate-gated channels. Cell 67(1), 11–19.

Spera, S. and Bax, A. (1991) Empirical correlation between protein backbone conformation and C_α and C_β ^{13}C nuclear magnetic resonance chemical shifts. J. Am. Chem. Soc. 113, 5490–5492.

Srinivasan, S., March, C.J. and Sudarsanam, S. (1993) An automated method for modeling proteins on known templates using distance geometry. Protein Sci. 2(2), 277–289.

States, D.J. and Karplus, M. (1987) A model for electrostatic effects in proteins. J. Mol. Biol. 197, 122–130.

Stewart, P.L., Fuller, S.D. and Burnett, R.M. (1993) Difference imaging of adenovirus: Bridging the resolution gap between X-ray crystallography and electron microscopy. EMBO J. 12(7), 2589–2599.

Still, W.C., Tempczyk, A., Hawley, R.C. and Hendrickson, T. (1990) Semianalytical treatment of solvation for molecular mechanics and dynamics. J. Am. Chem. Soc. 112, 6127–6129.

Stockman, B.J., Nirmala, N.R., Wagner, G., Delcamp, T.J., DeYarman, M.T. and Freisheim, J.H. (1992) Sequence-specific ^{1}H and ^{15}N resonance assignment for human dihydrofolate reductase in solution. Biochemistry 31, 218–229.

Suiko, M, Fernando, P.H., Sakakibara, Y., Nakajima, H., Liu, M.C., Abe, S. and Nakatsu, S. (1992) Posttranslational modification of protein by tyrosine sulfation: Active sulfate PAPS is the essential substrate for this modification. Nucleic Acids Symp. Ser. 183–184.

Swindells, M.B. (1994) Finding your fold (Commentary). Protein Eng. 7(1), 1–3.

Tanford, C. and Kirkwood, J.G. (1957) Theory of protein titration curves. I. General equations for impenetrable spheres. J. Am. Chem. Soc. 79, 5333–5339.

Tanford, C. and Roxby, R. (1972) Interpretation of protein titration curves. Application to lysozyme. Biochemistry 11, 2192–2198.

Teng, B., Burant, C.F. and Davidson, N.O. (1993) Molecular cloning of an apolipoprotein B messenger RNA editing protein. Science 260(5115), 1816–1819.

Topham, C.M., McLeod, A., Eisenmenger, F., Overington, J.P., Johnson, M.S. and Blundell, T.L. (1993) Fragment ranking in modelling of protein structure. Conformationally constrained environmental amino acid substitution tables. J. Mol. Biol. 229(1), 194–220.

Toyoshima, C., Sasabe, H. and Stokes, D.L. (1993) Three-dimensional cryo-electron microscopy of the calcium ion pump in the sarcoplasmic reticulum membrane. Nature 362(6419), 467–471.

Tramper, J., Vermuë, M.H., Beeftink, H.H. and van Stockar, U. (Eds.) (1992) Biocatalysis in Non-conventional Media. Proceedings of an International Symposium, Noordwijkerhout, The Netherlands, 26–29 April 1992. Elsevier, Amsterdam.

Tuchscherer, G., Servis, C., Corradin, G., Blum, U., Rivier, J. and Mutter, M. (1992) Total chemical synthesis, characterization, and immunological properties of an MHC class I model using the TASP concept for protein de novo design. Protein Sci. 1(10), 1377–1386.

Thunnissen, A.-M.W.H., Dijkstra, A.J., Kalk, K.H., Rozeboom, H.J., Engel, H., Keck, W. and Dijkstra, B.W. (1994) Doughnut-shaped structure of a bacterial muramidase revealed by X-ray crystallography. Nature 367, 750–753.

Ulrich, A.S., Heyn, M.P. and Watts, A. (1992) Structure determination of the cyclohexene ring of retinal in bacteriorhodopsin by solid-state deuterium NMR. Biochemistry 31, 10390–10399.

Unwin, N. (1993) Nicotinic acetylcholine receptor at 9 Å resolution. J. Mol. Biol. 229(4), 1101–1124.

Van Belle, D., Couplet, I., Prevost, M. and Wodak, S.J. (1987) Calculations of electrostatic properties in proteins. J. Mol. Biol. 198, 721–735.

van Tilbeurgh, H., Egloff, M.-P., Martinez, C., Rugani, N., Verger, R. and Cambillau, C. (1993) Interfacial activation of the lipase – procolipase complex by mixed micelles revealed by X-ray crystallography. Nature 363, 814–820.

van Tilbeurgh, H., Sarda, L., Verger, R. and Cambillau, C. (1992) Structure of the pancreatic lipase – colipase complex. Nature 359, 159–162.

Vriend, G., Sander, C. and Stouten, P.F.W. (1994) A novel search method for protein sequence – structure relations using property profiles. Protein Eng. 7(1), 23–29.

Wagner, G. (1990) NMR investigations of protein structure. Prog. NMR Spectrosc. 22, 101–139.

Wagner, G. (1993) Prospects for NMR of large proteins. J. Biomol. NMR 3, 375–385.

Wagner, G., Braun, W., Havel, T., Schaumann, T., Go, N. and Wüthrich, K. (1987) Protein structures in solution by nuclear magnetic resonance and distance geometry. J. Mol. Biol. 196, 611–639.

Warshel, A. and Levitt, M. (1976) Theoretical studies of enzymic reactions. J. Mol. Biol. 103, 227–249.

Warshel, A. and Russel, S.T. (1984) Calculation of electrostatic interactions in biological systems and in solution. Q. Rev. Biophys. 17, 283–422.

Warshel, A., Naray-Szabo, G., Sussman, F. and Hwang, J.-K. (1989) How do serine proteases really work? Biochemistry 28, 3629–3637.

Warwicker, J. and Watson, H.C. (1982) Calculation of the electric potential in the active site cleft due to α-helix dipoles. J. Mol. Biol. 157, 671–679.

Wendoloski, J.J. and Matthew, J.B. (1989) Molecular dynamics effects on protein electrostatics. Proteins Struct. Funct. Genet. 5, 313–321.

Wider, G., Macura, S., Kumar, A., Ernst, R.R. and Wüthrich, K. (1984) Homonuclear two-dimensional ^{1}H NMR of proteins. Experimental procedures. J. Magn. Reson. 56, 207–234.

Williamson, M.P. and Asakura, T. (1991) Calculation of chemical shifts of protons on alpha carbons in proteins. J. Magn. Reson. 94, 557–562.

Wilmanns, M. and Eisenberg, D. (1993) Three-dimensional profiles from residue-pair preferences: Identification of sequences with beta/alpha-barrel fold. Proc. Natl. Acad. Sci. USA 90(4), 1379–1383.

Winkler, F.K., D'Arcy, A. and Hunziker, W. (1990) Structure of human pancreatic lipase. Nature 343(6260), 771–774.

Wishart, D.S., Sykes, B.D. and Richards, F.M. (1992) The chemical shift index: A fast and simple method for the assignment of protein secondary structure through NMR spectroscopy. Biochemistry 31, 1647–1651.

Wishart, D.S., Willard, L. and Sykes, B.D. (1994) Vadar. In preparation.

Witkowski, A., Witkowska, H.E. and Smith, S. (1994) Reengineering the specificity of a serine active-site enzyme. Two active-site mutations convert a hydrolase to a transferase. J. Biol. Chem. 269(1), 379.

Woodcock, S., Mornon, J.P. and Henrissat, B. (1992) Detection of secondary structure elements in proteins by hydrophobic cluster analysis. Protein Eng. 5(7), 629–635.

Woolley, G.A. and Wallace, B.A. (1992) Model ion channels: gramicidin and alamethicin. J. Membr. Biol. 129, 109–136.

Wüthrich, K.(1986) NMR of Proteins and Nucleic Acids. John Wiley & Sons, New York.

Xu, M.-Q., Southworth, M.W., Mersha, F.B., Hornstra, L.J. and Perler, F.B. (1993) In vitro protein splicing of purified

precursor and the identification of a branched intermediate. Cell 75, 1371–1377.

Yang, A.-S., Gunner, M.R., Sampogna, R., Sharp, K. and Honig, B. (1993) On the calculation of pK_a's in proteins. Proteins Struct. Funct. Genet. 15, 252–265.

Yoshimura, S., Onozawa, T., Mizoguchi, J., Suemizu, H., Moriuchi, T. and Watanabe, K. (1990) Molecular cloning of cDNA coding for rat plasma glutathione peroxidase. Nucleic Acids Symp. Ser. 1990(22), 71–72.

Zauhar, R.J. and Morgan, R.S. (1985) A new method for computing the macromolecular electric potential. J. Mol. Biol. 186, 815–820.

Zhang, C.T. and Chou, K.C. (1992) An optimization approach to predicting protein structural class from amino acid composition. Protein Sci. 1(3), 401–408.

Zhou, G., Xu, X. and Zhang, C.T. (1992) A weighting method for predicting protein structural class from amino acid composition. Eur. J. Biochem. 210(3), 747–749.