14

DNA and RNA structure prediction

ERIC WESTHOF, PASCAL AUFFINGER, and CHRISTINE GASPIN

1. Introduction

An understanding of the functional mechanisms of a biological macromolecule requires the knowledge not only of its precise molecular organization in space but also of its internal dynamics. Molecular modelling attempts to construct the three-dimensional (3D) structure of a macromolecule on the basis of a mixture of theoretical and experimental data. Hence, prediction methods range from the most mathematically oriented ones, relying solely on computer algorithms, to the most pragmatical and operational one in which insights come alternatively from theory and experiment. Our contention is that modelling and simulation are most interesting in molecular biology when they possess a high predictive power.

Thus, we view modelling as a heuristic tool which should help in the rationalization of experimental observations but also, and most importantly, should suggest new relations between the various components of the modelled molecule. Without a 3D model, mutagenesis of a macromolecule will be, by necessity, somewhat random and, not always informative. In the absence of a 3D model able to organize the data at a higher level, mutagenesis experiments performed under such conditions will mainly confirm an available secondary structure (2D) of a RNA molecule. Such experiments can be useful, however, for bootstrapping a 3D structure which will serve as a framework for organizing existing data and suggesting new mutagenesis. Further, the history of structural discovery shows that there is no correlation between either accuracy or precision and predictive power. For example, molecular biology was born with the 1953 paper by Watson and Crick on the DNA double helix (1), but the structure, although accurate, was not precise by present standards.

The power of visualizing 3D relations is such that models need not always be detailed. On the contrary, extremely precise and detailed X-ray structures can be of no use for uncovering or understanding the function of a crystallized molecule without prior or further biochemical exploration and characterization. In the end, the validity and the accuracy of the model obtained

will depend on the nature of the experimental observations collected. However, a mathematical proof guaranteeing the correctness of the derived model is only possible with crystallographic methods (the Fourier theorem). Otherwise, the best that can be achieved is a network of evidence converging on the spatial contacts and relations embodied by a model.

The experimental observations used for deriving a 3D structure can be of quite different nature depending on the techniques employed and on the chemical nature of the macromolecule: from biophysical methods (partial X-ray diffraction data, NMR couplings, or NOEs, and other spectroscopic methods like UV, RAMAN, or circular dichroism), to biochemical approaches (chemical probing or enzymatic attack), and biological data (sequences, phylogenies). High-resolution X-ray crystallographic analysis (diffraction data at 1.5–1.0 Å resolution) yields a wealth of unequalled 3D information. However, this requires not only the crystallization of the macromolecule but also the solution to a phase problem. Generally, with biological macromolecules, the problem is compounded by their size and complexity. Besides, nucleic acids are very difficult to crystallize, since they are highly charged macromolecules which, in the case of RNA molecules, can undergo spontaneous cleavages. In addition, large, nucleic acids and especially RNAs, often exchange between various base pairings and foldings. Recently, NMR methods have proved their usefulness in this area. Chemical and enzymatic probing of nucleic acids in solution yields important information on the stability of the structures and on those bases protected from chemical or enzymatic attack. However, such experimental approaches will not reveal the nature of the interacting partners. Cross-linking experiments have the potential to give that information, but the cross-linking reactions take place in an assembly of molecules generally not all in the same state, and it is difficult to prove that the reactions occurred solely on functional molecules. Sequence data are extremely rich in potential 3D information, since they result from adaptative evolution over millions of years. Thus, if the function is identical and the sequences are sufficiently diverse, the noise level (or covariations resulting from contingencies) will be decreased by sequence comparisons. However, the extraction of 3D content from sequences is difficult and the method will strongly depend on the type of macromolecule under study. For example, self-splicing autocatalytic group I and group II introns, which require only water and ions to function, are more amenable to sequence comparisons than the catalytic RNase P RNA in ribonucleic particles which contains the history of its evolution with the tRNA substrate and with the protein co-factor.

The former experimental approaches (2–4) will not be discussed here. However, it should be kept in mind that the methods described in this chapter range from those in which the incorporation of experimental data is restricted to physical chemistry to those which use and exploit biological information. Molecular mechanics and dynamics belong to the first category.

RNA secondary structure prediction is simplified and on firmer ground with the incorporation of biological and chemical information, and successful RNA 3D modelling is best achieved on the basis of sequence comparisons and chemical probing.

2. Molecular mechanics and molecular dynamics methods

Molecular mechanics (MM) minimizes a particular energy function for a molecular system. The energy function contains steric and geometric terms as well as terms related to atomic interactions. A specific force-field is associated with a given energy function. Molecular dynamics (MD) simulations use similar force-fields and energy functions but, by integration of Newton's equation of motion, allow one to generate time-dependent trajectories of chemical or biochemical systems (5–7). These methods are usually used to add a dynamical perspective to systems for which time-dependent experimental knowledge is scarce and, most importantly, they are also used to process and refine crystallographic (8) or NMR data (*AMBER* (9), *Xplor* (10, 11)), or to calculate free energy differences between related systems by perturbation methods (12, 13). The advantages of MM are its easy implementation and short computing times. The main drawback is that the system might become locked in false minima which depend on possible inaccuracies resulting from the construction of the initial coordinate set or on the choice of the starting conformation. One way to relieve these undesirable effects is to minimize several starting conformations by varying one or more internal coordinates (torsion angles, for example). On the other hand, MD simulations, combined with energy minimizations, are well adapted to the sampling of the conformational space and the localization of local or global energy minima. As it is impossible to recommend, at the actual level of the technique, any definite protocol that one could follow in order to obtain physically meaningful MD simulations, we choose to discuss in this chapter general methodological details with, as guideline protocols, those that we apply in our laboratory on simulations of hydrated DNA and RNA fragments which include the aqueous environment and the counterions (14–16). Other details on simulations of nucleic acids can be found in two reviews (6, 17).

2.1 The potential energy function

The potential energy function, which describes in a simplified way the interactions between the atoms constituting the system, is central to the problem of molecular mechanics and molecular dynamics. A general form of this function, used in the *AMBER* MD package (9), is given by the equation:

$$E = \sum_{bonds} k_d(d-d_0)^2 + \sum_{angles} k_\theta(\theta-\theta_0)^2 + \sum_{dihedrals} \frac{V_n}{2}[1 + \cos(n\phi-\gamma)]$$

$$+ \sum_{nonbonded} \frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^{6}} + \sum_{nonbonded} \frac{q_i q_j}{\varepsilon r_{ij}}$$

where the first three terms represent the interactions between atoms separated by less than three bonds, the fourth and fifth term corresponding respectively to the van der Waals and electrostatic interactions occurring between non-bonded atoms. In the electrostatic term, the dielectric parameter ε is either a constant or a function of the distance between the charges.

In addition to the classical terms mentioned above, there is a great variety of additional terms describing for example 10–12 hydrogen bonds (18–20), or mixed terms which couple bond length and bond angle vibrations (21). Other specifics (like choice of parameters, of options, of functions, etc.) can be found in the *AMBER* (18, 19), *CHARMm* (20), *GROMOS* (22), or *OPLS* (23, 24) force-fields. It should be noted that the use of united atom force-fields, where the CH, CH_2, and CH_3 groups are represented by large hydrophobic atoms, compared with the all atom force-fields, is no longer justified, either for simulations *in vacuo* or for simulations taking into account a solvent environment, since the gain in computer time is not worth the approximations introduced in the system.

The choice of a set of partial atomic charges is of particular importance. Next to the classical ways of extracting charges from quantum mechanics calculations, charges derived from experiment were published (25) and recently tested in our laboratory on a simulation of the anticodon arm of $tRNA^{Asp}$. They were shown to give better agreement with known experimental structures than the standard *AMBER* set of charges. Other methods like multipole distributions, in which partial charges are no longer restricted to the atomic positions, have to be considered in the future to increase the accuracy of the electrostatic representation. Ultimately, with adequate computational power, a full electrostatic treatment taking into account the atomic polarizability will be necessary (7). A choice has also to be made concerning the water model to be used. Improving the classical SPC, TIP3P, or TIP4P models (26), the SPC/E model (27) is known to reproduce the diffusion coefficients of water and, therefore, should give more reliable time-dependent quantities.

The treatment of long-range electrostatic interactions (proportional to $1/r_{ij}$) is an issue of great concern. Because of computational limitations, it is very difficult to calculate electrostatic interactions up to a distance greater than a given cut-off value, usually 8–10 Å. This truncation method is not very

satisfactory and some authors have shown that they introduce non-negligible artefacts in the calculations (28, 29). To ameliorate the straight truncation of electrostatic forces, a wide range of switching and shifting functions has been employed and discussed (30, 31). Other approaches are possible, like the Ewald summation method used in simulations of a rigid DNA duplex with various counterions and co-ions (32) and of a rigid DNA triple helix in a 1.0 M NaCl aqueous solution (33).

2.2 Molecular dynamics simulation protocols

The following protocols are of course not unique and have to be adapted to each particular system and to the available computational means.

2.2.1 Construction of the system

All available experimental knowledge should be used in order to choose reasonable and interesting starting configurations. They can be extracted from the NDB (Nucleic Acid Data Base) (34) which contains most of the published crystallographic nucleic acid structures as well as structures derived from NMR experiments. Some of those structures are also contained in the PDB (Protein Data Bank) (35). Subsequently, the molecule needs to be solvated. This can be achieved at various levels of approximation. Partial solvation can be performed by putting a shell of water around the entire solute, or only around a site of particular interest (complexation or catalytic sites). There are different ways of constraining the solvent molecules located at the surface of the solvation shell, but some researchers let the water move freely in their simulations. We chose to use periodic boundary conditions which try to mimic an infinite system by replicating images of the simulation shell around the central box. However, the truncation distance used for computing long-range forces limit the range of the 'infinity' of the model.

Next, counterions are placed around the solute. Two methods are generally used for nucleic acids. The first consists in placing the ion along the bisector of the OPO angle at a distance of 4.5–6 Å from the phosphorus atom (9), and the other consists in replacing water molecules with the highest electrostatic potential by counterions until neutrality or the desired total charge is obtained (36). Various counterions have been used such as Na^+, K^+, NH_4^+, Ca^{2+}. To our knowledge, no simulations using the high structuring Mg^{2+} ion have been undertaken so far. The choice and positioning of counterions can be circumvented by reducing, according to the Manning theory of counterion condensation (17), the charges on the phosphate groups, and omitting explicit representation of the ions. However, this leads to values for the charges on the phosphate group below those of some polar atoms in the bases, and therefore alters considerably, and perhaps unrealistically, the water–phosphate interactions.

2.2.2 Equilibration and thermalization

In order to produce stable MD simulations, a good equilibration protocol which avoids early deformations of the solute originating from strong and unfavourable solute–solute, solute–solvent, and solvent–solvent interactions, is essential. Usually, at the beginning, a few hundred steps of energy minimization is used to relieve the main unfavourable constraints from the starting configuration, and, afterwards, the system is brought to equilibrium in several stages (*Figure 1*). First, the solvent alone is allowed to move around the fixed solute and counterions at constant temperature (300 K) and volume. Then, the constraints on the counterions are removed and 1 psec of dynamics at constant temperature and pressure (1 atm.) is performed at respectively 100, 200 K, followed by 5 psec at 300 K. Finally, the whole system is thermalized by a gradual increase of the temperature at each psec from 50 to 300 K by steps of 50 K. Subsequently, the heating step is followed by 5 psec of equilibration at 300 K (16).

Some authors have used constraints to maintain the base pairing of the starting structures during the equilibrium step, or even during the whole simulation. Our recent results proved that this is not always necessary. Breaking of base pairs can result from insufficient equilibration as well as from inaccurate force-field or simulation parameters.

2.2.3 Vacuum simulations

Simulations using no solvent have the advantage of being extremely fast. This allows one to conduct longer simulations and to sample more extensively the

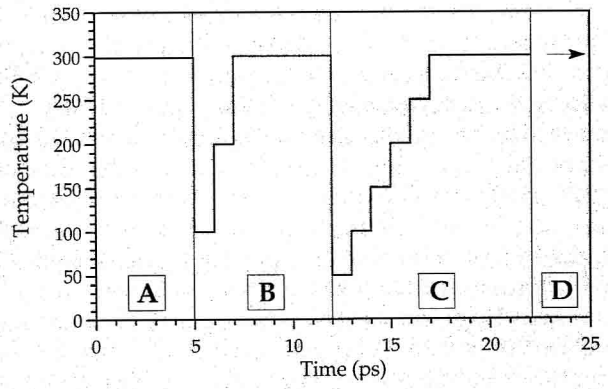


Figure 1. Equilibration protocol for a molecular dynamics simulation of a solvated nucleic acid with counterions. During part A, the solvent alone is allowed to move at 300 K; during part B, constraints are relieved from the counterions in steps of 100 K; during part C, no constraints are applied on the system but after cooling the system down to 50 K, it is warmed up in steps of 50 K; part D corresponds to the subsequent production phase at constant temperature and pressure.

configurational space. To compensate for the absence of solvent and counterions, various dielectric functions have been proposed, but they can give only approximate results since it is well known that specific local interactions with water are necessary to maintain the three-dimensional structure of nucleic acids (37).

2.3 Modelling large nucleic acids

For large nucleic acids, the all atom approach is no longer feasible. In order to be able to simulate such systems, Malhotra *et al.* (38) have developed models of varying resolution ranging from one pseudoatom per helix to one pseudoatom per nucleotide. This allows them to obtain useful but, consequently, much less precise information on the structure of these molecules.

2.4 Analysis of the trajectories

The analysis of the results of the calculations is the last but not the least important part of MD simulations. For nucleic acids, the 'Curves' (39) procedure for helical analysis has been used in a computer graphics utility called 'Dials and Windows' (40) which can monitor and display the time evolution of all the conformational and helical parameters in a DNA oligonucleotide.

3. Fine structure and the search for specific regions in DNA

DNA is not solely a storage medium for genetic information. Any sequence also contains control regions directing the binding of specific proteins as well as regions with static curvature or thermal lability. The prediction of the fine structure of DNA, i.e. the effects of base sequence on 3D structure, is the subject of an enormous literature (41). Here, we will refer more specifically to those methods which possess documented softwares. For small systems (up to 200 base pairs), molecular mechanics methods, as developed in programs such as *AMBER*, *GROMOS*, or *JUMNA* (42) have been used, especially in conjunction with NMR data. *JUMNA* is particularly well adapted to nucleic acids with helical periodicity, either DNA or RNA with between one and four strands in parallel or antiparallel orientations. The study of small systems either by MM (41) or MD (15), allows the extraction of the behaviour of more global parameters (like the average twist angle between two given base pairs or the average roll and tilt angles of a given base pair, i.e. the rotation about the long, respectively short, axis of a base pair, see *Figure 2*). Those parameters can then be inserted in programs using schematic and non-atomic representations of base pairs. Such programs are especially useful for visualizing the path of the helical axis as a function of intra- or interbase pair parameters. Four programs have been extensively used for the prediction

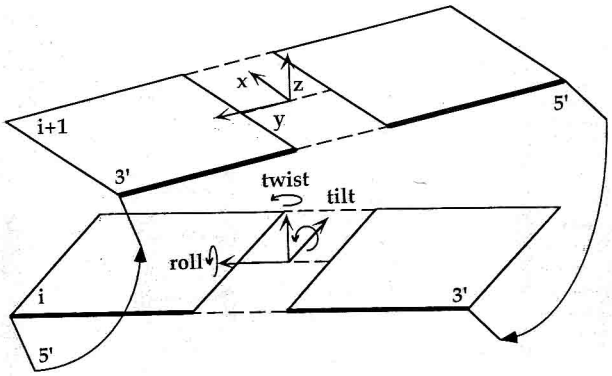


Figure 2. The six parameters relating a base pair to the next one in a double-stranded helix: the three translations along *x, y, z* (shift, slide, rise) and the three rotations about the *z* axis (twist angle or rotation angle between base pairs), the *y* axis (roll angle), and the *x* axis (tilt angle). For a complete discussion, see Dickerson *et al.* (113).

and display of bent DNA fragments. Bending results from curvature in the plane of the helical axis (controlled mainly by the roll angle) and torsion out of the plane (controlled by variations in the twist angle). The tilt angle is never large because of the resulting compression of the sugar-phosphate backbone. In two programs, *CURVATURE* (43) and that of De Santis *et al.* (44), a given set of those parameters or a mixture of them is used to compute the DNA path. In two other programs, *AUGUR* (45) and *DNA* (46), the user can choose among several sets of parameters or even introduce their own set.

4. RNA secondary structure prediction

Folded 3D RNA molecules are stabilized by a variety of interactions, the most prevalent of which are stacking and hydrogen bonding between bases on strands oriented in antiparallel directions. The 2D structure gives a subset of those interactions represented by Watson–Crick canonical (C–G, G–C, A–U, and U–A) and wobble (G–U and U–G) pairs of bases in double-stranded helices. Such a 2D fold provides an important constraint for determining the 3D structure of RNA molecules (47, 48). Therefore, the determination of the 2D structure is an essential step in the study of the structure–function relationships. Another task associated with RNA 2D structures concerns the automatic identification of specific RNAs in genomic DNA sequences (49, 50).

The determination of a 2D structure results generally from the combination of several approaches, each one using specific knowledge depending on

the presence of a set of homologous sequences or of only a single sequence. This section is mainly devoted to the theoretical description of current methods of RNA 2D folding and the associated available programs which are in use today.

4.1 Representation

A 2D fold can be represented on a circle graph where the N nucleotides of the sequence are represented as vertices (dots) and are connected by edges representing the phosphodiester bonds between consecutive nucleotides (along the circle) and hydrogen bonds between bases (across the circle). A valid 2D structure is usually defined as a structure for which the graph contains only edges which do not cross each other. Such a graph is a planar graph. In a more conventional representation, computed with programs such as *Squiggle* (51), *LoopViewer* (52), and *Rnasearch* (53), where bonds are represented as edges of nearly the same size, the folding gives rise to characteristic secondary structural elements which are usually divided into six different types (*Figure 3*): helices, single-stranded regions, bulges, internal loops, hairpin loops, and multibranched loops.

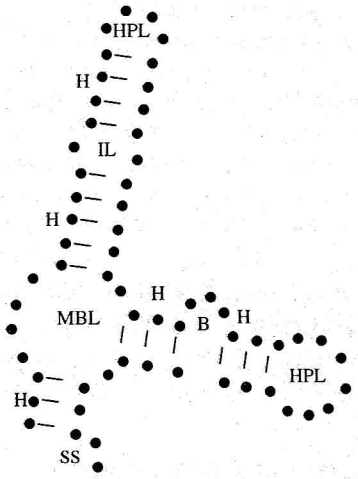


Figure 3. Secondary structural elements. HPL represents a hairpin loop which is formed when an RNA strand folds back on itself. IL represents an internal loop. At least one base is unpaired on each strand of the loop separating two paired regions. A mismatch is a special type of internal loop for which only one nucleotide on each strand is not Watson–Crick paired. B represents a bulge. A bulge has unpaired nucleotides on only one strand. The other strand has uninterrupted base pairing. H represents a helix. A helix is a region of consecutive pairs of bases. MBL represents a multibranched loop or junction. A multibranched loop occurs when double-stranded regions separated by any number of unpaired nucleotides come together. SS represents an unpaired region.

4.2 Data necessary for folding RNA molecules

When a set of homologous sequences (homologous sequences have common ancestry and function) is available, one can search for compensatory base changes which maintain base-paired helices with the help of an available alignment. When only one sequence is available or when RNAs are not conserved among a sufficiently diverse set of organisms, theoretical models of predictions have to be associated with experiments. The related knowledge is based on a set of constraints, the thermodynamic model and the available experimental data on the molecule.

4.2.1 The constraints

Models of prediction generally include the following restrictions on the folding of an RNA into a secondary structure:

(a) Pair restriction forbids all the non-canonical pairings allowing only A–U (two hydrogen bonds), G–C (three hydrogen bonds), and G–U (two hydrogen bonds) pairs.

(b) Uniqueness restriction allows at most one pairing for each base.

(c) Pseudoknot restriction forbids pseudoknots. Two base pairs numbered (i, j) and (k, p) form a pseudoknot if $i < k < j < p$ or $k < i < p < j$. Pseudoknots (*Figure 4*) result from Watson–Crick base pairing involving a stretch of bases in a loop between paired strands and a distal single-stranded region (which could belong itself to a hairpin loop or a bulge). Thus a pseudoknot is akin to a special case of 3D base pairing rather than a structural 2D element. Because efficient programs are essentially based on the ability to decompose a structure into substructures, which is not possible if pseudoknots exist, pseudoknots are usually taken care of in a second step.

(d) Stereochemical restriction requires that at least three ribonucleotides separate two paired strands of ribonucleotides because the chemical linkages cannot stretch beyond a certain distance.

(e) Length restriction affects the length of a helix (the number of base pairs) and allows only helices with a length greater than a given value (usually two).

These restrictions lead to the determination of what is commonly called a valid secondary structure although the restrictions are not always well-founded. In bulges, non-Watson–Crick pairing, such as U–U, A–A, and A–G pairs, are often observed (e.g. 5S rRNA) (2). Also, the existence of unusually stable tetraloops (54), like -GNRA- or -UNCG-, with a non-Watson–Crick pair between G and A (or U and G) shows that hairpin loops can be made with only two unpaired bases (55). Finally, pseudoknots are also extremely frequent in structured RNAs (e.g. group I introns) (56) as in control regions of mRNAs and lead to ambiguities in the 2D definition (57).

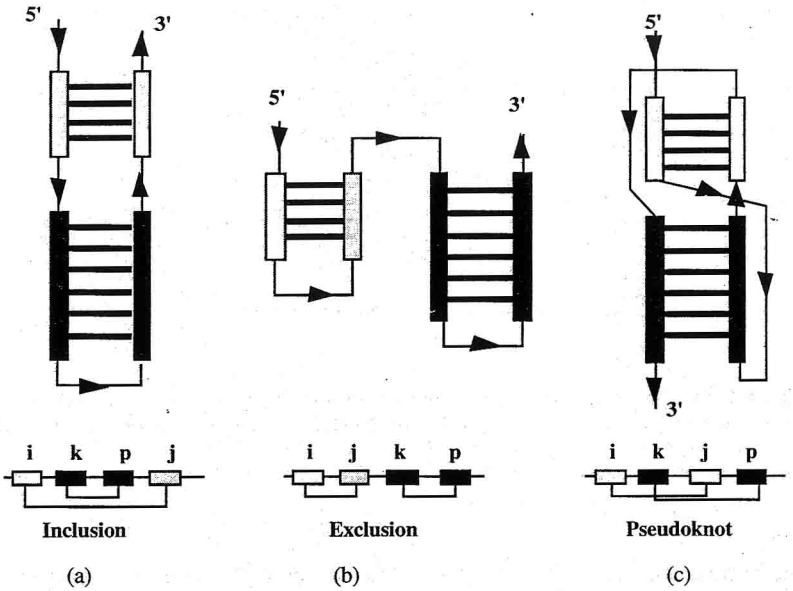


Figure 4. The three possible relationships between two double-stranded helical regions: in (a) and (b) two standard helices and in (c) a pseudoknot. In (a), the paired strands (*i,j*) are included between the paired strands (*k,p*) giving a long hairpin interrupted by an internal loop. In (b), the paired strands (*i,j*) and the paired strands (*k,p*) form two adjacent hairpins. In (c) the paired elements alternate along the sequence ($i < k < j < p$), leading to a pseudoknot structure (see ref. 56 for a detailed description of the 3D structures of pseudoknots and of their functions).

4.2.2 The thermodynamic model

The thermodynamic stability of structural elements has been studied to evaluate their probability of formation. These values, computed from experiments on short sequences of nucleotides, give an estimation of the stabilizing free energy of base stacking as well as the destabilizing free energy of single strands. Based on such a set of parameters, several thermodynamic models exist. From the simplified values of Tinoco (58) to Ninio's sophisticated model (59), dedicated to the 5S rRNAs and tRNAs, the most used model nowadays is that of Turner (60). However, it has not been possible to obtain experimental values for each elementary motif, which makes the thermodynamic model rough and uncomplete. For example, until recently, all the loops were considered as destabilizing elements whereas some tetraloops have recently been shown to be very stable (54). Most of the 2D folding programs now take into account the complete model of Turner as well as the parameters associated with the tetraloops (61, 62).

Moreover, the thermodynamic model assumes the Tinoco–Uhlenbeck

postulate which states that the free energy of the whole structure is the sum of the free energies of its secondary structural elements. The assumption that the energy of a position in the folded structure is affected only by its nearest neighbours is certainly not correct but the additivity assumption works well and is essential to all prediction algorithms.

4.2.3 Available experimental data

Various enzymes and chemicals are available for probing the solution structure of RNA thus giving detailed data at the nucleotide level. Thus, a real map of the single- or double-stranded regions in the molecule can be established. The mechanisms of action of the probes, the limitations of the technique, and the methods for detection of cuts or modifications are described elsewhere (2, 3). With these data, a number of potential structural elements can be eliminated from consideration in the calculation of folding.

When several mutually exclusive secondary structures exist, site-directed mutagenesis can be used to test for compensatory base changes in the potential helices. Because experiments are time consuming and because precise probing of each nucleotide is difficult, theoretical models of prediction try to incorporate, whenever possible, available data on the studied molecule. The incorporation of this information in 2D folding programs is actually the only way to produce a correct structure.

4.3 Methods of prediction

4.3.1 Sequence comparisons

Comparative analysis of nucleic acid sequences has been widely used for the detection and evaluation of similarities and evolutionary relationships. With RNA molecules, sequence alignments and RNA 2D prediction are intimately related. Comparative analysis is based on the biological paradigm that macromolecules are the product of their historical evolution and that functionally homologous sequences will adopt similar structures. The sequences are first aligned and then searched for compensatory base pair changes. If, during evolution, a base has been modified in a strand of a potential helix (mutation), then this modification must have been compensated on the complementary strand in order to maintain the structure. The presence of several compensatory changes (two or more) in a potential helix allows one to assert the existence of the helix in the structure. Several secondary structure models have been generated by using comparative analysis: tRNA (63), 5S RNA (64), 16S RNA (65), 23S RNA (66), RNase P RNA (4), group I and group II self-splicing introns (67, 68). The method requires that the molecules compared must be sufficiently different to provide enough instances of sequence variations with which to test pairing possibilities but that the molecules do not differ so much that homologous regions cannot be aligned with confidence.

i. Alignment

The objective is to juxtapose related sequences so that homologous residues in each sequence occupy the same column in the alignment (*Figure 5*). Since the 1970 program of Needleman and Wunsch (69), programs to align more than two sequences have been put forth using different strategies including reduction of the problem to three sequences (70), application to closely related sequences (71), the help of a predetermined evolutionary tree (72), the search for common subsequences (73), or the selection of the best

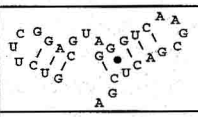
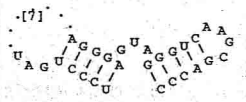
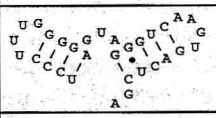
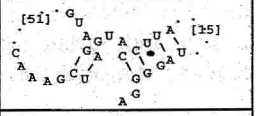
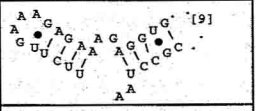
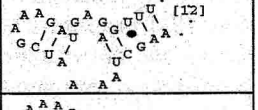
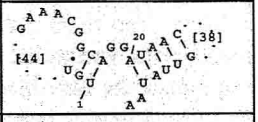
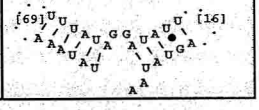
SG	N°	P7.1 P7.1' P7.2 P7.2'	
IA2	73	GUCUUCG------ GACGUAGGGUCAAGCGACUCGA	
	74	UCCCUGAU[7]AGGGAGUAGGGUCAAGCGACCC GA	
	75	UCCCUUUG---- GGGAGUAGGGUCAAGUGACUCGA	
	76	UCGAAAC[51]GUAGAGUACCUUA[15]UAGGGG A	
IA3	77	UUCUU--GAAAGAGAAAG-AGGUG[9]CGCCUAA	
	78	AAUC---GAAA-GAUGAG-AGUUU[12]AAGCUAA	
	79	UGU[44]GAAACGGCAGG-AUAAC[38]GUUAUAA	
	80	UAUAAA[69]UUUAUAGG-AUAUU[16]AGUAUAA	

Figure 5. Part of the alignment of group I introns corresponding to the structural elements P7.1 and P7.2 of two subgroups IA2 and IA3. (Extracted from the appendix of ref. 48.) The paired sequences are underlined. The numbers correspond to the sequence numbering of ref. 48.

pairwise alignments to gradually align sequences by using an order of incorporation of sequences into the final alignment (*PileUp* (51), *CLUSTAL* (74), and *MultAlin* (75)). Other programs dedicated to the alignment of RNA sequences allow the user to manipulate interactively the proposed alignment (*DCSE* (76), *ALIGNOS* (77)). They offer functions dedicated to secondary structures as well as an interactive environment for manipulating the alignment.

Other recent and interesting programs automatically reconsider the alignment by taking into account new sequences and pre-existing knowledge of the secondary structure (78, 79). Indeed, with the growing number of sequences, specific RNA databases are created and new sequences have to be quickly added to structured databases of homologous RNA molecules. In such databases, it is very desirable that sequences be aligned in accordance with the conserved secondary structural features. Because, in an alignment, optimal structural elements can be misaligned, the program *RNAlign* makes it possible to align a group of aligned sequences with a new sequence, using positions of high sequence conservation and common secondary structures the group as a guide for determining the secondary structure of the new sequence. Thus, *RNAlign* does not suppose that the related sequences are correctly aligned but instead reconsiders the alignment. *RNAlign* was used to build a structured database of RNA from the large ribosomal subunit. The other method of multiple alignment (79), which differs from all those described above, uses stochastic context-free grammars (80) to build a statistical model during, rather than after, the process of alignment and folding. Such an approach was applied to the multiple alignment of tRNA.

ii. Comparative analysis

Given an ordered sequence alignment, comparative analysis can begin. Most computerized approaches to comparative analysis are based on the number of varying positions in base pairs of Watson–Crick helices (81–83). Han and Kim (83) propose a very simple algorithm that builds a covariation matrix where one can visualize, by different characters and for each possible pair of positions, a complementary base change (for each sequence, the base in column i can form a Watson–Crick pair with the base in column j), an exact match (no variation in both columns i and j), a wobble pair (in most sequences the base in column i can form a G–U pair with a base in column j), an inexact pair (a base i does not form a pair with a base in column j for each sequence and the number of pairs is greater than a threshold value) or a mismatch (a base i does not form a pair with a base in column j for each sequence and the number of pairs is lower than a threshold value). In this matrix, possible helices (diagonals of characters) are combined in order to compute valid common secondary structures.

However, all these programs rely on an available alignment which may not be unique, especially when the sequences are highly divergent in primary struc-

ture. Moreover, in an alignment optimal for classical scores, the preserved secondary structural elements can be misaligned. Therefore comparative analysis programs such as those presented above have to be used with caution.

4.3.2 Energy minimization

The usual criterion for computing the RNA secondary structure of a single sequence is to minimize the free energy of the folded molecule. Several types of algorithms, among which are *Fold* (84) and *CRUSOE* (85) have been used to find the optimal secondary structure. These methods have been described extensively (60, 84, 86) and will not be described here. Instead, we will describe the main principles of each one and, whenever they exist, the extensions that have been realized in order to compute more appropriate secondary structures.

i. Dynamic programming approaches

The most commonly used algorithm is based on dynamic programming, first used by Nussinov and Jacobson (87). The main advantage of this type of algorithm is speed and thus the ability to fold large molecules. However, they compute only one optimal structure. These algorithms work by first computing optimal structures for fragments of five nucleotides then extending the fragments one nucleotide at a time in both directions until the fragment becomes the whole sequence.

Instead of computing the minimum free energy structure, the partition function of all possible structures and the pairing probability for every possible pair can be calculated, using a dynamic programming algorithm described by McCaskill (88). This program, which is available in the Vienna package (62), allows one to process base pair probabilities through a postscript dotplot where each base pairing probability is represented by a square of corresponding value in the upper part of the matrix. The lower part of the matrix contains the minimum free energy structure according to Zuker's method. In these programs, the temperature at which the base pairings are computed can be varied, as can the choice of the set of energy parameters related to the various elementary structural elements.

ii. Combinatorial approaches

The second type of algorithm usually called a 'combinatorial' approach, works in two steps. It first generates all the possible helices that can be formed from the sequence and then combines them into valid structures (85, 89, 90). This approach, however, is generally limited to molecules with less than 200 bases because of the exponential number of possible combinations.

4.3.3 Extensions of dynamic programming approaches

These algorithms are ultimately limited by our partial understanding of the parameters necessary for the calculation of the free energy. Accordingly,

optimally folded structures may not represent the actual base pairing relationships found in the RNA molecule, either because several folded structures with very similar free energies are possible, or because other cellular elements stabilize active RNA structures that otherwise would be thermodynamically less stable. A partial solution is to extend folding programs to allow for the calculation of a range of possible structures that take into account a given set of biochemical data.

Suboptimal folding is the process of determining a set of possible folded structures that have very similar free energy minima but different foldings. Combinatorial approaches can easily compute a set of suboptimal structures. For the case of dynamic programming, several approaches have been developed. The most popular of these is that of Zuker (91), but there are others (92, 93). In the extension developed by Zuker (91), which is an adaptation of the optimal folding method, the result of the suboptimal folding is a series of structures that have similar free energy minima. It is based on the observation that a fold containing a pair (b_i, b_j) divides the structure into two parts: a folding of the included fragment b_i to b_j and a folding of the excluded fragment from b_j to b_i. The two quantities $V(i,j)$ and $V(j,i)$ are computed, $V(i,j)$ representing the minimum folding energy of the included fragment and $V(j,i)$ representing the minimum folding energy of the excluded fragment. In order to compute suboptimal secondary structures, the strategy consists in identifying all bases pairs for which $V(i,j) + V(j,i)$ is close to E_{min}, the energy of an optimal folding of the sequence from 1 to N. In this extension, a P-optimal base pair is defined so as to be contained in at least one folding within P percent of the minimum free energy. Optimal and suboptimal foldings can be generated either automatically or by selecting a base pair. In the first case, optimal and suboptimal foldings are sorted by energy. In the second case, optimal or suboptimal foldings contain the chosen base pair.

In the original package (94), analysis of suboptimal structures is aided by two ways of visualizing the RNA fold: the energy dotplot and a plot of the number of possible different base pairs versus nucleotide position in the sequence (P-num graphs). The program is able to consider various constraints on the folding such as locations of single-stranded sites, double-stranded sites or known helices. It is also possible to force regions to pair together, one region to pair anywhere, one region to be single-stranded or two regions not to pair together. The energy parameters used are those of Turner (60) with additional values for tetraloops.

4.3.4 Interactive computer assisted approaches

Approaches which provide an environment in which the experimentalist can participate in the computational folding of the RNA molecule are called 'interactive' approaches. The strength of interactive approaches lies in their ability to test different structural constraints without modification of the folding program. Structures can thus be continuously modified according to new

biochemical information and the user is free to compare biochemical constraints according to intuition. This type of approach (95) is supported by a computer program which allows:

- the examination of as many of the possible substructures as desired
- the use of filtering to incorporate information on pairing length, pairing and stacking energies, experimental data, user assumptions
- the incorporation of related sequences
- user selection and evaluation.

A dotplot matrix, in which the sequence is compared to its reverse complement, allows the visualization of potential helices for selection by the user. A secondary structure is not calculated with this approach. Instead, helices are chosen, then analysed with respect to two criteria such as the energy of the helix and chemical/enzymatic data. Cedergren *et al.* (96) have incorporated the same approach into an RNA folding editor. The program, called *RNASE*, consists of two main units: a helix editor and a structure editor. The user may select desired helices in the secondary or tertiary structure among a list of computed helices. These helices are verified for overlap before being combined into a secondary structure.

The interactive approach we have developed (97) incorporates restrictions from the length of helices and the available data before the step selection and is able to take into account all the usual constraints. In this way, only the possible pairings can be chosen during the selection step. Moreover, the formalism used and the associated algorithms allow one to consider other types of constraints as well as secondary structures with pseudoknots, by adding or removing appropriate constraints. In the selection step, selected elements are not helices but individual pairs of bases. Moreover, energetic criteria encoding the free energy of the molecule is not necessarily taken into account in the search procedure. However, such a criterion can be considered through a selection probability matrix of pairing like that proposed by McCaskill (88) in which the selected pairs become the most probable pairs in accordance with the thermodynamic criteria.

4.3.5 Sequential folding

This type of method relies on the simulation of the folding process (98–102). In these methods, the folding is considered to be a stepwise process where intermediate structures evolve into the native one by subsequent addition of preferred stems. Generally, the programs start to fold the sequence by adding the most stable stems assuming that these are kinetically favoured and act as nucleation centres for local RNA folding. In one method (98), a competition between helices is performed by using random structure generation. The consideration of folding during synthesis is performed by calculating several cycles of folding determination for each incomplete RNA sequence and

increasing the sequence after each cycle. In these programs, pseudoknots are allowed to be nucleation centres.

4.4 Limits

4.4.1 Complexity of algorithms

The time complexity of optimal folding methods increases at least approximately with the cube of the length of the sequence, even with a simplification hypothesis, which constitutes a potential limitation. One way to calculate the folded structure of a large RNA is to fold consecutive subregions of the molecule (61), keeping in mind that dynamic programming methods tend to favour pairing of the 5′- and 3′-extremities.

4.4.2 Significance of folded structures

Without experimental data, assessment of the significance of a folded structure is very difficult. Several strategies have been used. For example, alternative foldings can be calculated for a sequence by first using suboptimal folding methods or by varying parameters. Results are then compared and those foldings in which motifs appear systematically may be considered as significant. It is also possible to refold the molecule in successively overlapping pieces, to compare the motifs that arise, and to keep as significant only those that are reproducible (99). A third method is to fold several random sequences that have the same base composition and compare the folding energies (100).

Moreover, computed minimal energy structures may not be biologically relevant. The problem does not lie merely in the uncompleteness of the thermodynamic parameter sets, the naivety of simple additive models or the fact that input thermodynamic values were derived under conditions that may not truly mimic *in vivo* situations. The ultimate difficulty is rather that many natural RNAs are likely to require helpers (proteins or other RNAs) which control their folding into biologically active forms.

5. RNA tertiary structure construction

Construction of the tertiary structure of an RNA molecule always starts from a given secondary structure. Insights about tertiary contacts can be gained through chemical modifications (which give the relative importance of specific atomic positions) or probing (some protections cannot be explained by the 2D structure), by cross-linking experiments (which directly indicate the partners, assuming a single conformer in solution) and, most efficiently, by careful sequence comparisons (48). The approaches divide themselves into those which rely on mathematical objectivity and automation to those which exploit partial and potentially biased human decisions. In the first category is included the distance geometry method (103) although there are problems

choosing the correct chiralities and for avoiding knots in the structures. Another method, *YAMMP* (104), exploits a pseudoatom approach with either one pseudoatom per helix or one pseudoatom per nucleotide. The use of spherical pseudoatoms, however, leads to a loss in the asymmetry of the RNA fragments and, most importantly, all fine interactions which control RNA folding are not modelled. A third approach is based on a constraint satisfaction algorithm. The program, *MC-SYM* (47, 105) searches conformational space such that, for a given set of input constraints (secondary pairings, tertiary pairs, distances), all possible models are produced. With this methodology, Major *et al.* (106) managed, for a tRNA sequence, to generate 26 solutions which displayed the broad features of canonical tRNA structure. Our own approach involves an extensive use of known structures. The framework of those structures is held in a database which is used by the program *FRAGMENT* for inserting the appropriate sequence (106, 107). The fragments produced are then assembled manually on a graphics screen using any modelling software (*FRODO*, *INSIGHT*, *PRO-EXPLORE*). The resulting structure is then refined by restrained least-squares minimization programs (*NUCLIN/NUCLSQ*) (108). Molecular mechanics or molecular dynamics could also be employed at this stage. The manipulations on the screen imply some human judgements which depend on the knowledge of 3D structure and the personal bias of the modeller. However, the human mind can quickly exclude sets of solutions and take into account experimental data. The solvent accessibilities of the final model can be easily computed (e.g. *ACCESS*) (109) to validate the structure against experimental reactivities of specific positions to chemical reagents.

6. Conclusions

Table 1 is a compilation of the programs discussed in the present chapter together with the address of the contacting author or distributor. The programs are classified according to the main topics of the chapter. Unfortunately, the programs are often dedicated to some specific machine or system and it is not always convenient to go back and forth between the requested or produced input/output files. At the present time, there is no comprehensive package able to deal with the various aspects of nucleic acid modelling. The development of such packages is in dire need.

Acknowledgements

This research was supported by the GIP-GREG (92H0906), the Ministère de l'éducation nationale, the GDR-1029 'Informatique et Génomes', and the CM2AO program of ORGANIBIO (P.A.).

Table 1. Overview of the programs discussed in the chapter with the address of author or distributor

Program	Key words	Source/reference
Alignments and comparative analysis		
ALIGNOS	Alignment editor	(77)
DCSE	Alignment editor	(76)
RNAlign	Reconsideration of alignment—databases	(78) fcorpet@toulouse.inra.fr
Klinger and Brutlag	Comparative analysis	(82)
Han and Kim	Comparative analysis	(83)
COVARIATION	Comparative analysis	(110) FTP site : iubio.bio.indiana.edu
2D folding programs		
CRUSOE	(Sub)optimal—combinatorial	(85) mgouy@evomol.univ-lyon1.fr
RNASE	Editor, interactive, computer assisted	(96) Montréal University, Canada
McCaskill	Partition function	(62) FTP site: ftp.itc.univie.ac.at
Abrahams	Sequential folding	(99)
MFOLD	(Sub)optimal—dynamic programming	(61) [a]
2D and 3D drawing programs		
Drawna	Automatic 3D ribbon drawings	(40) westhof@ibmc.u-strasbg.fr
Rnasearch	Automatic drawing without overlapping	(53) gaspin@toulouse.inra.fr,
Squiggle	Automatic drawing	[a]
LoopViewer	Automatic drawing	Indiana University, Bloomington, USA Don.Gilbert@IUBio.Bio.Indiana.Edu FTP site : 129.79.224.25
Molecular mechanics and molecular dynamics packages		
AMBER	Free energy calculations, structure refinements,...	(18, 19) amber@cgl.ucsf.edu
CHARMm	Free energy calculations, structure refinements,...	(20)
GROMOS	Free energy calculations, structure refinement	(36)
Xplor	Molecular dynamics and structure refinement	(10) Yale University, New Haven, CT, USA
Quanta/Charmm	Interactive graphics based on *CHARMm*	Polygen Corporation, Waltham, MA, USA
Insight/Discover	Interactive graphics and molecular mechanics	BIOSYM Technologies, San Diego, CA, USA
Macromodel	Interactive graphics and molecular mechanics	(111) Columbia University, New York
PROSIMULATE PROEXPLORE	Interactive graphics based on *GROMOS*	Oxford Molecular, The Magdalen Centre, Oxford OX4 4GA, UK
FRODO	Interactive graphics, construction and manipulations of 3D structures	(112)
TURBO FRODO	Interactive graphics, construction and manipulation of 3D structures	Biographics, Marseille, France turbo@lccmb.cnrs-mrs.fr
JUMNA	Junction minimizations of nucleic acids (and nucleic acid–ligand complexes)	(39) IBPC, 13 rue Pierre et Marie Curie, Paris, France
YAMMP	MM on large nucleic acids	(104)
MC-SYM	Conformational search program	(105) major@tremblant.nlm.nih.gov
Nuclin/Nuclsq	Least squares structure refinement	(108) westhof@ibmc.u-strasbg.fr

[a] Genetics Computer Group, Inc., University Research Park, 575 Science Drive, Suite B, Madison, Wisconsin 53711- Help@GCG.Com (51).

References

1. Watson, J. D. and Crick, F. H. C. (1953). *Nature*, **171**, 737.
2. Ehresmann, B., Ehresmann, C., Romby, P., Mougel, M., Baudin, F., Westhof, E., *et al.* (1990). In *The ribosome, structure, function, and evolution* (ed. W. E. Hills, A. Dahlberg, R. A. Garett, P. B. Moore, D. Schlessinger, and J. R. Arner), pp. 148–59. American Society for Microbiology, Washington D.C.
3. Krol, A. and Carbon, P. (1989). In *Methods in enzymology* (ed. J. N. A. Simon and M. I. Simon), Vol. 180, p. 212. Academic Press, London.
4. Woese, C. R. and Pace, N. R. (1993). In *The RNA world* (ed. R. F. Gesteland and J.F. Atkins), p. 91. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.
5. Allen, M. P. and Tildesley, D. J. (ed.) (1987). *Computer simulation of liquids*. Clarendon Press, Oxford.
6. McCammon, J. A. and Harvey, S. C. (ed.) (1987). *Dynamics of proteins and nucleic acids*. Cambridge University Press, Cambridge.
7. Van Gunsteren, W. F. and Berendsen, H. J. C. (1990). *Angew. Chem. Int. Ed. Engl.*, **29**, 992.
8. Gros, P., Fujinaga, M., Mattevi, A., Vellieux, F. M. D., Van Gunsteren, W. G., and Hol, J. (1989). In *Molecular simulation and protein crystallography (Proceedings of the Joint CCP4/CCP5 Study Weekend) SERC* (ed. J. Goodfellow, K. Henrik, and R. Hubbard), p. 1. Daresbury Laboratory, UK.
9. Pearlman, D. A., Case, D. A., Caldwell, J. C., Seibel, G. L., Singh, U. C., Weiner, P., *et al.* (1991). *AMBER 4.0*. University of California, San Francisco.
10. Brünger, A. T. (1990) XPLOR. Yale University, New Haven, CT.
11. Brünger, A. T. (1990). In *Molecular dynamics: applications in molecular biology* (ed. J. M. Goodfellow), pp. 137–78. Macmillan Press, London.
12. Beveridge, D. L. and DiCapua, F. M. (1989). *Annu. Rev. Biophys. Biophys. Chem.*, **18**, 431.
13. McCammon, J. A. (1991). *Curr. Opin. Struct. Biol.*, **1**, 196.
14. Fritsch, V. and Westhof, E. (1991). *J. Am. Chem. Soc.*, **113**, 8271.
15. Brahms, S., Fritsch, V., Brahms, J. G., and Westhof, E. (1992). *J. Mol. Biol.*, **223**, 455.
16. Westhof, E., Rubin-Carrez, C., and Fritsch, V. (1995). In *Computer modelling in molecular biology* (ed. J. M. Goodfellow), pp. 103–31. VCH, NY.
17. Beveridge, D. L., Swaminathan, S., Ravishanker, G., Whithka, J. M., Srinivasan, J., Prevost, C., *et al.* (1993). In *Water and biological macromolecules* (ed. E. Westhof), Vol. 17, pp. 165–225. Macmillan Press Ltd., London.
18. Weiner, S. J., Kollman, P., Case, D. A., Singh, C. U., Ghio, C., Alagona, G., *et al.* (1984). *J. Am. Chem. Soc.*, **106**, 765.
19. Weiner, S. J., Kollman, P. A., Nguyen, D.T., and Case, D. A. (1986). *J. Comput. Chem.*, **7**, 230.
20. Brooks, B. R., Bruccoleri, R. E., Olafson, B. D., States, D. J., Swaminathan, S., and Karplus, M. (1983). *J. Comput. Chem.*, **4**, 187.
21. Dauber-Osguthorpe, P., Roberts, V. A., Osguthorpe, D. J., Wolff, J., Genest, M., and Hagler, A. T. (1988). *Proteins*, **4**, 31.
22. Hermans, J., Berendsen, H. J. C., Van Gunsteren, W. F., and Postma, J. P. M. (1984). *Biopolymers*, **23**, 1513.

23. Jorgensen, W. L. and Tirado-Rives, J. (1988). *J. Am. Chem. Soc.*, **110**, 1657.
24. Pranata, J., Wierschke, S. G., and Jorgensen, W. L. (1991). *J. Am. Chem. Soc.*, **113**, 2810.
25. Pearlman, D. A. and Kim, S. H. (1990). *J. Mol. Biol.*, **211**, 171.
26. Jorgensen, W. L., Chandrasekhar, J., and Madura, J. D. (1983). *J. Chem. Phys.*, **79**, 926.
27. Berendsen, H. J. C., Grigera, J. R., and Straatsma, T. P. (1987). *J. Phys. Chem.*, **97**, 6269.
28. Schreiber, H. and Steinhauser, O. (1992). *Biochemistry*, **31**, 5856.
29. Schreiber, H. and Steinhauser, O. (1992). *J. Chem. Phys.*, **168**, 75.
30. Smith, P. E. and Pettitt, B. M. (1991). *J. Chem. Phys.*, **95**, 8430.
31. Kitson, D. H., Avbelj, F., Moult, J., Nguyen, D. T., Mertz, J. E., Hadzi, D., *et al.* (1993). *Proc. Natl. Acad. Sci. USA*, **90**, 8920.
32. Forester, T. R. and McDonald, I. R. (1991). *Mol. Phys.*, **72**, 643.
33. Mohan, V., Smith, P. E., and Pettitt, B. M. (1993). *J. Phys. Chem.*, **97**, 12984.
34. Berman, H. M., Olson, W. K., Beveridge, D. L., Westbrook, J., Gelbin, A., Demeny, T., *et al.* (1992). *Biophys. J.*, **63**, 751.
35. Bernstein, F. C., Koetzle, T. F., Williams, G. J. B., Meyer, E. F., Brice, M. D., Rodgers, J. R., *et al.* (1977). *J. Mol. Biol.*, **112**, 537.
36. Van Gunsteren, W. F. and Berendsen, H. J. C. (1987). *Groningen Molecular Simulation (GROMOS)*, Library Manual Biomos, Groningen.
37. Westhof, E. and Beveridge, D. L. (1990). *Water Sci. Rev.*, **5**, 24.
38. Malhotra, A., Gabb, H. A. G., and Harvey, S. C. (1993). *Curr. Opin. Struct. Biol.*, **3**, 241.
39. Lavery, R. and Sklenar, H. (1988). *J. Biomol. Struct. Dynam.*, **6**, 63.
40. Ravishanker, G., Swaminathan, S., Beveridge, D. L., Lavery, R., and Sklenar, H. (1989). *J. Biomol. Struct. Dynam.*, **6**, 669.
41. Lavery, R. (1994). In *Advances in computational biology* (ed. O. V. Hugo), Vol. 1, p. 69. JAI Press Inc., Greenwich, Connecticut.
42. Lavery, R. (1988). In *Structure and expression* (ed. W. K. Olson, R. H. Sarma, M. H. Sarma, and M. Sundaralingam), Vol. 3, p. 191. Adenine Press, New York.
43. Shpigelman, E. S., Trifonov, E. N., and Bolshoy, A. (1993). *Comput. Appl. Biosci.*, **9**, 435.
44. De Santis, P., Fuà, M., Palleschi, A., and Savino, M. (1993). *Biophys. Chem.*, **46**, 193.
45. Tan, R. K. Z., Prabhakaran, M., Tung, C. S., and Harvey, S. C. (1988). *Comput. Appl. Biosci.*, **4**, 147.
46. Treger, M. and Westhof, E. (1987). *J. Mol. Graph.*, **5**, 178.
47. Major, F., Turcotte, M., Gautheret, D., Lapalme, G., Fillion, E., and Cedergren, R. (1991). *Science*, **253**, 1255.
48. Michel, F. and Westhof, E. (1990). *J. Mol. Biol.*, **216**, 585.
49. Lisacek, F., Diaz, Y., and Michel, F. (1994). *J. Mol. Biol.*, **235**, 1206.
50. Fichant, G. A. and Burks, C. (1991). *J. Mol. Biol.*, **220**, 659.
51. Genetics Computer Group (1991). *Program Manual for the GCG Package, Version 7.* Madison, WI.
52. Gilbert, D. (1990). *LoopViewer Package*, Bloomington.
53. Muller, G., Gaspin, C., Etienne, A., and Westhof, E. (1993). *Comput. Appl. Biosci.*, **9**, 551.
54. Antao, V. P. and Tinoco, I. (1992). *Nucleic Acids Res.*, **20**, 819.

55. Westhof, E., Romby, P., Romaniuk, P., Ebel, J.-P., Ehresmann, C., and Ehresmann, B. (1989). *J. Mol. Biol.*, **207**, 417.
56. Westhof, E. and Jaeger, L. (1993). *Curr. Opin. Struct. Biol.*, **2**, 327.
57. Westhof, E. and Michel, F. (1994). In *RNA–protein interactions: frontiers in molecular biology*. pp. 25–51. IRL Press, Oxford.
58. Tinoco, I., Uhlenbeck, O., and Levine, M. (1971). *Nature*, **230**, 362.
59. Papanicolaou, C., Gouy, M., and Ninio, J. (1984). *Nucleic Acids Res.*, **12**, 31.
60. Turner, D. H. and Sugimoto, N. (1988). *Annu. Rev. Biophys. Biophys. Chem.*, **17**, 167.
61. Jaeger, J. A., Turner, D. H., and Zuker, M. (1989). *Proc. Natl. Acad. Sci. USA*, **86**, 7706.
62. Hofacker, I., Fontana, W., Stadler, P. F., Bonhoeffer, L. S., Tacker, M., and Schuster, P. (1994). *Monatshefte für Chemie.* **125**, 167.
63. Holley, R. W., Apgar, J., Everett, G. A., Madison, J. T., Marquisee, M., Merrill, S. H., Penswick, S. H., and Zamir, J. R. (1965). *Science*, **147**, 1462.
64. Fox, G. E. and Woese, C. R. (1975). *Nature*, **256**, 505.
65. Woese, C. R., Magrum, L. J., Gupta, R., Siegel, R. B., Stahl, D. A., Kop, J., *et al.* (1980). *Nucleic Acids Res.*, **8**, 2275.
66. Noller, H. F., Kop, J., Wheaton, V., Brosius, J., Gutell, R., Kopylov, A. M., *et al.* (1981). *Nucleic Acids Res.*, **9**, 6167.
67. Davies, R. W., Waring, R. B., Ray, J. A., Brown, T. A., and Scazzocchio, C. (1982). *Nature*, **300**, 719.
68. Michel, F., Jacquier, A., and Dujon, B. (1982). *Biochimie*, **64**, 867.
69. Needleman, S. B. and Wunsch, C. D. (1970). *J. Mol. Biol.*, **48**, 443.
70. Murata, M., Richardson, J. S., and Sussman, J. L. (1985). *Proc. Natl. Acad. Sci. USA*, **82**, 3073.
71. Bains, W. (1989). *Comput. Appl. Biosci.*, **5**, 51.
72. Sankoff, R. J. and Cedergren, G. L. (1976). *J. Mol. Evol.*, **7**, 133.
73. Martinez, H. (1988). *Nucleic Acids Res.*, **16**, 1683.
74. Higgins, D. G., Bleasby, A. J., and Fuchs, R. (1992). *Comput. Appl. Biosci.*, **8**, 189.
75. Corpet, F. (1988). *Nucleic Acids Res.*, **16**, 10881.
76. De Rijk, P. and De Wachter, R. (1993). *Comput. Appl. Biosci.*, **9**, 735.
77. Neurath, H. and Wolters, J. (1992). *Bioinformatics*, **1**, 22.
78. Corpet, F. and Michot, B. (1994). *Comput. Appl. Biosci.*, **10**, 389.
79. Sakakibara, Y., Brown, M., Underwood, R.C., Mian, I.S., and Haussler, D. (1994). In *Proceedings of the 27th Hawaii International Conference on System Sciences*, Hawaii.
80. Searls, D. B. (1992). In *Artificial intelligence and molecular biology* (ed. L. Hunters), p. 47. AAAI Press/The MIT Press, Cambridge, MA.
81. Winker, S., Overbeek, R., Woese, C. R., Olsen, G. J., and Pfluger, N. (1990). *Comput. Appl. Biosci.*, **6**, 365.
82. Klinger, T. M. and Brutlag, (1993). In *Proceedings of ISMB93* (ed. L. Hunter), p. 225. Bethesda, Maryland.
83. Han, K. and Kim, H.-J. (1993). *Nucleic Acids Res.*, **21**, 1251.
84. Zuker, M. and Stiegler, P. (1981). *Nucleic Acids Res.*, **9**, 133.
85. Gouy, M. (1987). In *Nucleic acid and protein sequence analysis: a practical approach* (ed. M. Bishop and C. J. Rawlings), p. 259. IRL Press, Oxford.

86. Zuker, M. (1989). In *Methods in enzymology* (ed. J. E. Dahlberg. and J. N. Abelson), Vol. 180, pp. 262–88. Academic Press, London.
87. Nussinov, R. and Jacobson, A. B. (1980). *Proc. Natl. Acad. Sci. USA*, **77**, 6309.
88. McCaskill, J.S. (1990). *Biopolymers*, **29**, 1105.
89. Pipas, J. M. and McMahon, J. E. (1975). *Proc. Natl. Acad. Sci. USA*, **72**, 2017.
90. Studnicka, G. M., Rahn, G. M., Cummings, I. W., and Salser, W. A. (1978). *Nucleic Acids Res.*, **5**, 3365.
91. Zuker, M. (1989). *Science*, **244**, 48.
92. Yamamoto, K. and Yoshikura, H. (1985). *Comput. Appl. Biosci.*, **1**, 89.
93. Williams, A. L. and Tinoco, I. (1986). *Nucleic Acids Res.*, **14**, 299.
94. Jaeger, J. A., Turner, D. H., and Zuker, M. (1990). In *Methods in enzymology* (ed. R. F. Doolittle), Vol. 183, p. 281. Academic Press, London.
95. Auron, P. E., Rindone, W. P., Vary, C. P. H., Celentano, J. J., and Vournakis, J. N. (1982). *Nucleic Acids Res.*, **10**, 403.
96. Cedergren, R., Gautheret, D., Lapalme, G., and Major, F. (1988). *Comput. Appl. Biosci.*, **4**, 143.
97. Gaspin, C. and Westhof, E. (1995). *J. Mol. Biol.*, **254**, 163.
98. Gultyaev, A. P. (1991). *Nucleic Acids Res.*, **19**, 2489.
99. Abrahams, J. P., Van Den Berg, M., Van Batenburg, E., and Pleij, C. (1990). *Nucleic Acids Res.*, **18**, 3035.
100. Martinez, H. M. (1984). *Nucleic Acids Res.*, **12**, 323.
101. Le, S.-Y., Chen, J.-H., Currey, K. M., and Maizel, J. V. (1988). *Comput. Appl. Biosci.*, **4**, 153.
102. Le, S.-Y. and Maizel, J. V. (1989). *J. Theoret. Biol.*, **138**, 495.
103. Hubbard, J. M. and Hearst, J. E. (1991). *Biochemistry*, **30**, 5458.
104. Malhotra, A., Tan, R. K. Z., and Harvey, C. (1990). *Proc. Natl. Acad. Sci. USA*, **87**, 1950.
105. Major, F., Gautheret, D., and Cedergren, R. (1993). *Proc. Natl. Acad. Sci. USA*, **90**, 9408.
106. Westhof, E., Romby, P., Ehresmann, C., and Ehresmann, B. (1990). In *Theoretical biochemistry, and molecular biophysics* (ed. D. L. Beveridge and R. Lavery), Vol. 1, p. 399. Adenine Press, New York.
107. Westhof, E. (1993). *J. Mol. Struct. (Theochem)*, **286**, 203.
108. Westhof, E., Dumas, P., and Moras, D. (1985). *J. Mol. Biol.*, **184**, 119.
109. Richmond, T. J. (1984). *J. Mol. Biol.*, **178**, 63.
110. Brown, J. W. (1991). *Comput. Appl. Biosci.*, **7**, 391.
111. Mohamadi, F., Richards, N. G. J., Guida, W. C., Liskamp, R., Lipton, M., Caufield, C., *et al.* (1990). *J. Comput. Chem.*, **11**, 440.
112. Jones, T. J. (1978). In *Computational chemistry* (ed. D. Sayre), p. 303. Oxford University Press.
113. Dickerson, R. E., Bansal, M., Calladine, C. R., Diekmann, S., Hunter, W. N., Kennard, O., *et al.* (1989). *J. Mol. Biol.*, **205**, 627.

DNA and Protein Sequence Analysis

A Practical Approach

Edited by

M. J. BISHOP

UK HGMP Resource Centre,
Hinxton, Cambridge CB10 1SB, UK

and

C. J. RAWLINGS

SmithKline Beecham Pharmaceuticals
New Frontiers Science Park
Third Avenue, Harlow, Essex CM19 5AW, UK

IRL PRESS
at
OXFORD UNIVERSITY PRESS
Oxford New York Tokyo