CABIOS

Vol 7 no 3 1991
Pages 347 – 352

Inferring consensus structure from nucleic acid sequences

David K.Y.Chiu and Ted Kolodziejczak

Abstract

This paper presents an unsupervised inference method for determining the higher-order structure from sequence data. The method is general, but in this paper it is applied to nucleic acid sequences in determining the secondary (2-D) and tertiary (3-D) structure of the macromolecule. The method evaluates position–position interdependence of the sequence using an information measure known as expected mutual information. The expected mutual information is calculated for each pair of positions and the chi-square test is used to screen statistically significant position pairs. In the calculation of expected mutual information, an unbiased probability estimator is used to overcome the problem associated with zero observation in conserved sites. A selection criterion based on known structural constraints of the strongest interdependent position pairs is applied yielding position pairs most indicative of secondary and tertiary interactions. The method has been tested using tRNA and 5S rRNA sequences with very good results.

Introduction

The secondary (2-D) and tertiary (3-D) structures of a macromolecule play an important role in determining the interaction and functions of a macromolecule in biological systems. Current knowledge is adequate to correctly predict about 70% of the secondary structure from a single sequence (Jaeger *et al.*, 1990). Even though the higher-order structure of only a portion of the sequences is known, homologous sequences are often assumed to conform to a common structure. For example, because the cloverleaf model is agreed to be the most favorable configuration for tRNAs, the proposed folding structure for a newly sequenced tRNA may be accepted or refuted depending on whether it conforms to that common structure. Hence the common structure of homologous sequences is as important as the macromolecular structure of an individual sequence. We call the estimated common structure inferred from an ensemble of homologous sequence the consensus structure.

In nucleic acid sequences, prediction of higher-order structure using computational means involves searching for a folding configuration of maximum base pairing and/or minimum free energy (Benedetti *et al.*, 1989; Tinoco *et al.*, 1971; Waterman and Smith, 1978; Zuker and Stiegler, 1981; Gouy, 1987). This involves an exhaustive or heuristic search of all possible configurations for an optimal structure. Given a sequence with N nucleotides, the problem is analogous to finding an optimal graph with N nodes in satisfying certain constraints. Zuker and Sankoff (1985) estimated that when the nucleotides occur randomly with equal probability, the expected number of valid structures is greater than 1.8^N. Thus a sequence of 100 nucleotides has about 3×10^{25} possibilities and a computer evaluating 1000 structures per second would take 10^{15} years to find the structure. The very large number of possibilities makes this a difficult task and most algorithms, except for dynamic programming algorithms when the loop energies are sufficiently simple, consider only a portion of the configuration space (for examples see Gouy, 1987; Waterman, 1984; Zuker and Sankoff, 1984). Prediction of tertiary structure is often labor intensive and very difficult. Most existing computational methods can only be used to infer the secondary structure of a macromolecule.

Inferring higher-order structure from an ensemble of homologous sequences makes use of the redundancy of the sequences to conform to a consensus structure (for examples in phylogenetic studies). In principle, an ensemble of homologous sequences contains more information of the common structure than a single sequence alone. Instead of searching for a favorable folding configuration, our proposed method searches for the most favorable bonding positions in an aligned ensemble of sequences. As a result, the search space is drastically reduced to evaluate N^2 position-pairings. The method is based on the observation that base types in bonding positions of homologous sequences are highly correlated (Olsen, 1988; Zuker and Sankoff, 1984; Zuker and Stiegler, 1981). In addition, tertiary interaction is observed to exhibit base pairing fidelity similar to that seen in secondary structure (Haselman *et al.*, 1988, 1989). Our method evaluates position-pairings using both statistical and structural constraints. Thus the position-pairing detected is more informative than that using either statistical or structural constraints alone. From the detected position-pairings, secondary and tertiary structures of the macromolecules are determined.

Department of Computing and Information Science, University of Guelph, Guelph, Ontario, Canada

In the past, Haselman *et al.* (1988), Garnier *et al.* (1978), Gibrat *et al.* (1987) and Chiu *et al.* (1990, 1991) have evaluated the statistical information on the structure in nucleic acid and protein data. In Haselman's method, the covariation index is designed to indicate the extent of interaction. However, the reduction of the matrix in his method may delete useful information in the evaluation. This is because it relies on a dominant base pairing such as the Watson–Crick type and fails to consider position pairs that are either highly conserved or exhibit a high proportion of 'wobble' pairing. Garnier's method requires a supervised labeling scheme with the training data. Neither Haselman's nor Garnier's method have considered the secondary and tertiary structural constraints of the macromolecular structure. In Chiu *et al.* (1990, 1991), the structure is not described explicitly.

System and methods

The algorithms proposed in this paper were implemented in C on an IBM PS2/70 computer running under the DOS operating system. Nucleic acid sequences used are in the GenBank database.

Algorithms

In our proposed method, we calculate an information measure known as the expected mutual information between all possible position pairs for evaluating statistical interdependency. The information measure provides a quantitative measure of the interactive strength in a position–position pairing. Position pairs with significant statistical interdependence are identified and the ones with the strongest interdependency are selected. From the set of position pairs with the strongest interdependency, position pairs are classified based on whether they satisfy certain structural constraints using a circular representation extended from the Nussinov graph (Nussinov *et al.*, 1978). The secondary and tertiary structure of the macromolecule is then inferred and separated from the selected subset of position pairs.

Screening of position–position interactions

Let us describe our method formally. Consider a genetic sequence $x_1,x_2,x_3, \ldots, x_N$ where x_k, $1 \leq k \leq N$, can assume one of four different nucleotide types, denoted as $\{a_u | u = 1, \ldots, 4\}$ for the four nucleotides: A (adenine), G (guanine), C (cytosine), U (uracil) and N is the length of the sequence. Given an ensemble of m sequences, an alignment process is applied either manually or using a computational method (for examples see Chan *et al.*, 1991; Waterman, 1986). This involves the insertion of gaps where necessary to compensate for the missing nucleotides in the sequences. The length of sequences is thus increased to a common length n. The aligned set of m sequences of length n may be represented by the $m \times n$ pattern matrix $M = \{x_{ij}; i = 1, \ldots, m, j = 1, \ldots, n\}$, or

$$M = \begin{bmatrix} x_{11} & x_{12} & x_{13} & \ldots & x_{1n} \\ x_{21} & x_{22} & x_{23} & \ldots & x_{2n} \\ . & & & & . \\ . & & & & . \\ x_{m1} & x_{m2} & x_{m3} & \ldots & x_{mn} \end{bmatrix} \quad (1)$$

Let a variable X_j denote the j^{th} column of M where the outcome of X_j is the nucleotide type at the j^{th} position for a particular sequence. Consider the variable pair X_j and X_k. In general, not all the variable pairs are statistically interdependent. In our analysis, we initially select the statistically significant pairs using a test based on the expected mutual information or the rate of information transmission (Ash, 1965). It is expressed as:

$$I(X_j,X_k) = H(X_j) + H(X_k) - H(X_j,X_k) \quad (2)$$

where $H(X_j)$ is the estimated composition entropy or Shannon's entropy for the variable X_j and $H(X_j,X_k)$ is the interdependence entropy for the variable pair X_j and X_k (Ash, 1965).

$H(X_j)$ is defined as:

$$H(X_j) = -\sum_{u=1}^{4} P(X_j = a_u) \ln P(X_j = a_u) \quad (3)$$

and $H(X_j,X_k)$ is defined as:

$$H(X_j,X_k) = -\sum_{u=1}^{4}\sum_{v=1}^{4} P(X_j = a_u, X_k = a_v) \ln P(X_j = a_u, X_k = a_v) \quad (4)$$

Using equations (3) and (4), we can rewrite (2) as:

$$I(X_j,X_k) = \sum_{u=1}^{4}\sum_{v=1}^{4} P(X_j = a_u, X_k = a_v) \ln \frac{P(X_j = a_u, X_k = a_v)}{P(X_j = a_u)\,P(X_k = a_v)} \quad (5)$$

where $P(X_j = a_u)$ denotes the probability of a nucleotide type a_u occurring in X_j, and $P(X_j = a_u, X_k = a_u)$ denotes the probability of the joint occurrence of a_u and a_v in X_j and X_k respectively.

It has been shown that the expected mutual information has an asymptotic chi-square distribution (Wong and Liu, 1975; Wong and Chiu, 1987). Since each of the position variables can assume one of the four nucleotide types, ignoring any gaps, the degrees of freedom is equal to $(4 - 1) \times (4 - 1)$, or 9. The test for statistical interdependence becomes as follows. Given j and k, then X_j and X_k are statistically interdependent

if $I(X_j,X_k) \geq \chi^2/2m$, where m is the number of samples and χ^2 is the tabulated chi-square value with 9 degrees of freedom at a presumed significance level. The asymptotic chi-square distribution is true on the null hypothesis that the variables are independent. Thus a high score deviates from independence between the variables and refutes the assumption. The set of position pairs which have statistically significant interaction can be denoted by a set of ordered pairs:

$$S_1 = \{(j,k) | I(X_j,X_k) \geq \chi^2/2m;\ 1 \leq j,k \leq n\} \quad (6)$$

Unbiased probability estimator

When calculating the probability estimates on variables X_j and X_k based on frequency counts of the sampling data, we consider only those samples without a gap represented for either X_j or X_k. For highly conserved positions where some nucleotide type is not observed in the ensemble, instead of using the maximum likelihood estimation which assigns zero probability to this event of zero observation, an unbiased probability estimation originally proposed by Pascal (Mortimer, 1979; Wong and Chiu, 1987) is used. This is particularly useful when the sample size is very small. The unbiased probability estimator is defined in the following form:

$$P(X_j = a_u) = \frac{\{\text{number of } (X_j = a_u) \text{ observed}\} + 1}{\{\text{total number of sequences}\} + 4} \quad (7)$$

where a_u is a nucleotide type under consideration.

The unbiased probability estimator can be intuitively thought of as indicating an adjustment term in the estimation if four more samples were to be observed, then one of the samples would be the event a_u, assuming all nucleotide types are equally likely *a priori*. The unbiased probability estimator is equal to the maximum likelihood estimator if the sample size is infinitely large.

Using unbiased probability estimation, the probability of $(X_j = a_u, X_k = a_v)$ is estimated as:

$$P(X_j = a_u, X_k = a_v) = \frac{\{\text{number of } (X_j = a_u, X_k = a_v)\} + 1}{\{\text{total number of sequences}\} + 16} \quad (8)$$

The numbers 4 and 16 in the denominator of the estimator refer to the different possible nucleotide types in a position and between positions respectively.

The effect of using the unbiased probability estimator in the calculation of expected mutal information can be illustrated in the following example.

Example: Let us consider two positions j, k in an ensemble of 60 aligned sequences. Assume the two positions are completely conserved, i.e. say the nucleotide type observed for the positions X_j and X_k are A and G respectively. There are 16 different combinations of possible nucleotide types for the two positions and the expected mutual information calculated using the unbiased probability estimator is as follows:

$$\begin{aligned} I(X_jK_k) &= (61/76)\ln\{(61/76)/\{(61/64)(61/64)\}\} + \\ &\quad 6(1/76)\ln\{(1/76)/\{(1/64)(61/64)\}\} + \\ &\quad 9(1/76)\ln\{(1/76)/\{(1/64)(1/64)\}\} \\ &= 0.363 \end{aligned}$$

Note that the expected mutual information is reasonably high and statistically significant ($0.363 \geq \chi^2_{(0\ 99.9)}/2m = 0.181$). By comparison, the expected mutual information using the maximum likelihood estimation will yield a value of zero if the probability of unobserved nucleotide type is estimated as zero. Thus, the unbiased estimator will be more useful in handling cases with conserved positions. The calculated expected mutual information is then used in the statistical test in selecting the set S_1 in equation (6).

Grouping of position pairings

After identifying the set S_1, we select those position pairs which have the highest expected mutual information, or the most indicative of direct interaction in the ensemble. The rationale is that all indirect interaction will have a smaller expected mutual information. To perform this, we consider each position j and a corresponding position k such that the expected mutual information $I(X_j,X_k)$ is maximum. That is, a subset S_2 is chosen from S_1 as follows:

$$S_2 = \{(j,k) \mid I(X_j,X_k) = \max_{k'} I(X_j,X_{k'});\ j = 1, \ldots, n\} \subseteq S_1 \quad (9)$$

S_2 contains both the secondary and tertiary interactions. Other forms of interactions such as strong functional relationships may also be included.

To analyze the interactions further to include structural information, we have adopted the circular representation extended from Nussinov *et al.* (1978). Positions of the RNA sequence ensemble are numbered sequentially from the 5′ end and placed equidistant to one another along the circumference of a circle. Interactions between positions in the aligned sequences are then represented by chords joining pairs of positions. Note that the chords represent interactions of the aligned positions in the ensemble rather than of the positions of a single sequence as in the Nussinov graph (Nussinov *et al.*, 1978). This representation has the geometric property that helices generally appear as two or more parallel chords. Since only the secondary interactions exhibit this characteristic, we can identify interactions as secondary or tertiary.

Using the circular representation, two chords (j, k) and (j',k') are parallel (e.g. Figure 1) if they satisfy $j - j' = k' - k$, where $j < k$, and $j' < k'$. From this we introduce the following definition.

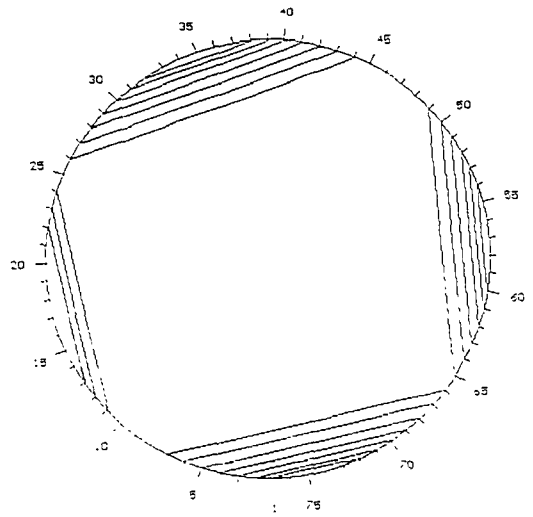

Fig. 1. Circular representation of secondary interactions inferred from 131 manually aligned tRNA sequences

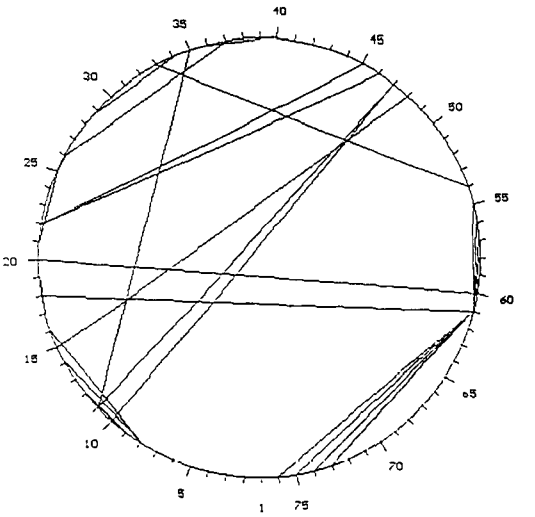

Fig. 2. Circular representation of tertiary interactions inferred from 131 manually aligned tRNA sequences.

Definition. Two chords (j,k) and (j',k') are D_d parallel neighbors if $j - j' = k' - k$, and $|j - j'| = d + 1 = |k - k'|$, where $d \in \{0,1,2,3, \ldots,\}$, and $j < k, j' < k'$.

Note that D_0 parallel neighbors represent two adjacent and parallel chords while D_1 parallel neighbors are two parallel chords separated by a gap. To introduce the consideration of structural constraints into the detection process, a position pair (j,k) is detected to be a member of a helical region if the following criteria are met: (i) a helix has a minimum length of two chords, with possible non-adjacent gaps within the helix, that is, a position pair (j,k) must have a D_0 or D_1 parallel neighbor (j',k'); (ii) the loops closing a helical region must consist of at least three bases, that is, a position pair (j,k) must satisfy $(k - j) > 3$ and its neighbor (j',k') must satisfy $(k' - j') > 3$.

The above criteria based on the idea of structural planarity can be used to differentiate position pairs most indicative of secondary interactions from other interactions. We select a subset S_3 from S_2 as follows. The set of S_3 contains those position pairs in S_2 having D_0 or D_1 parallel neighbors, or

$$S_3 \backslash S_2 \tag{10}$$

and

$$S_3 = \{(j,k) | j - j' = k' - k, |j - j'| \leq 2, |k - k'| \leq 2, (k - j) > 3, (k' - j') > 3; (j,k), (j',k') \in S_2\} \tag{11}$$

The set S_4 contains those members of S_2 not in S_3, or

$$S_4 = S_2 \backslash S_3 \tag{12}$$

Note that from the definition of S_3, the position pairs in S_3 are those with strongest statistical interdependency having at least one parallel chord that is at a distance of at most one gap. Thus, they are most likely to form a helical region corresponding to part of the secondary structure of a macromolecule. We consider those in S_4 as indicative of position pairs involved in other types of interactions. In our experiments, many of these interactions are tertiary interactions.

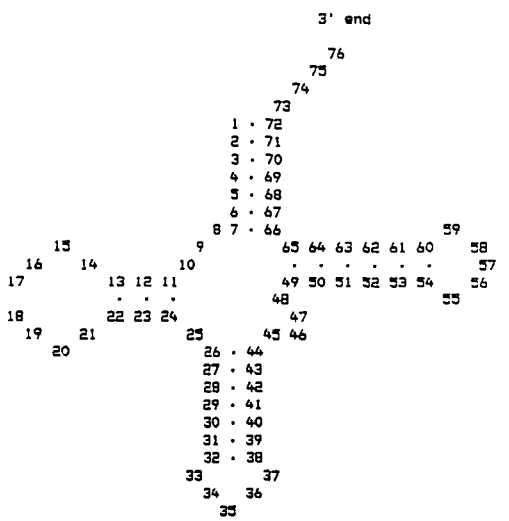

Fig. 3. Consensus secondary structure for the ensemble of 131 manually aligned tRNA sequences. Inferred position–position interactions are marked

Examples and discussion

Experiment 1: ensemble of 131 aligned tRNA sequences

We have applied our method to a set of 131 tRNA sequences of similar length chosen from the GenBank genetic sequence database (release 55.0). Although the set contains both initiator and elongator tRNAs, it is believed that the similarities in their structure allow us to include them so as to increase the size

The set of position pairs inferred as secondary interactions

(4,139), (5,138), (6,137), (7,136), (8,135), (9,134), (10,133), (23,72), (24,71), (25,70), (26,69), (38,57), (39,56), (40,55), (95,112), (96,111), (97,110), (99,108).

The set of position pairs inferred as tertiary interactions

(17,61), (19,61), (27,61), (30,61), (33,63), (34,36), (35,62), (37,61), (43,61), (44,49), (45,61), (46,61), (47,61), (50,61), (51,61), (52,61), (53,61), (54,61), (58,61), (61,64), (61,65), (61,90), (61,93), (61,114), (63,66), (81,82), (82,86), (90,91), (92,115), (99,109), (132,135).

Fig. 4. Positional pairings inferred from 34 5S rRNA sequences.

of the ensemble and use our method to represent the probabilistic variation in the data. We aligned the sequences manually to a common length of 76 by introducing as few gaps as possible to compensate for apparent insertions and deletions in different sequences. At most, two gaps are inserted for each sequence at the variable segment position of the *D* stem and the variable loop.

For the manually aligned set of sequences, we apply the proposed method and generate circular representations of position–position interactions (see Figures 1 and 2). Most of the known secondary and some tertiary interactions have been identified correctly as compared to the structure of *yeast* $tRNA^{Phe}$ (Grosjean *et al.*, 1982). The results are consistent with the cloverleaf model of tRNA. Figure 2 shows the secondary interactions and reveals four distinct base-paired regions corresponding to the helical regions of the secondary structure. The shortest detected helix consists of three stacked pairs and the longest one consists of seven stacked pairs. Included in the set of secondary interactions is the tertiary interaction of 13●22. Also detected are the tertiary interactions 26●44 and 32●38 suggested by an X-ray crystallographic study in Sussman and Kim (1976).

The known tertiary interactions of 8●14, 15●48, 22●46 have been correctly identified by our method. Some additional interactions were also found and it is possible that some of the pairs such as 33●54, 29●34, 11●35, 36●39 and 26●37 may be the result of functional relationships since they involve positions in the anti-codon region.

The tertiary interaction of 54●58, although statistically significant, is missed by our selection process due to the presence of the more favorable interaction of 54●60. When we examine the statistically significant interactions for these positions, we find that position pair 54●58 was a 'second best' in the expected mutual information value and that position 54 is involved in both the secondary and tertiary interaction.

The interaction of 14●21 is not selected because of the selection of 8●14. Similarly, the secondary interaction 10●25 is found to have a relatively lower statistically significant expected mutual information value and is not selected. In these cases, the 'second best' interaction is more relevant.

The inferred consensus secondary structure for the 131 tRNA sequence ensemble is shown in Figure 3. Positions are numbered sequentially beginning from the 5′ end. Interactions

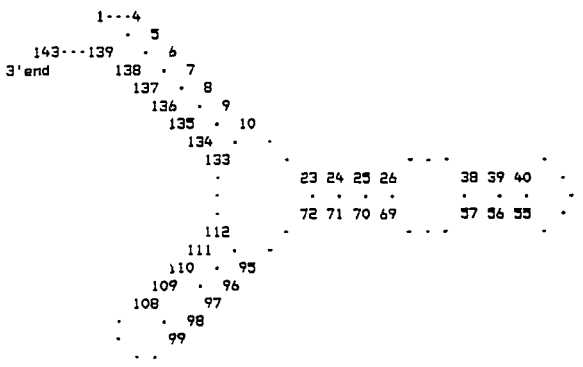

Fig. 5. Consensus secondary structure for the ensemble of 34 5S rRNA sequences showing inferred position–position interactions. Range of unpaired positions is indicated by '---'. The position index is based on the aligned sequences from Waterman (1988).

detected using our method are marked accordingly. The secondary and tertiary structure conform well to the known structure of *yeast* $tRNA^{Phe}$.

We have also performed an experiment on this set of sequences aligned using Waterman's algorithm for multi-sequence alignment (Waterman, 1986). However, the result is less stable depending on the value of parameters in the alignment algorithm. Despite the possibility of error in the alignment process, this experiment illustrates the usefulness of the method which can generate both a secondary and a tertiary structure testable by laboratory experiments.

Experiment 2: 34 aligned 5S rRNA sequences

In this experiment, an ensemble of 34 5S rRNA sequences is aligned on four consensus helices as described in Waterman (1988). Figure 4 shows the representation of the higher-order position–position interactions inferred using our method.

The inferred secondary interactions given as a set of position pairs in Figure 4 are used to construct consensus structure of the ensemble (see Figure 5). The diagram shows secondary base pairings as inferred using our method. The inferred position pairs are found to be in agreement with the consensus helices proposed by Waterman (1988). However, our method finds that some of the pairings within the consensus helices proposed by Waterman do not have statistically significant interaction. These are 1●119, 9●111, 10●110, 18●65, 23●60, 31●51, 79●97, 80●96, 80●95 and 85●91.

Conclusion

This paper has presented a method to infer the secondary and tertiary structure from an ensemble of aligned homologous sequences. The inferred higher-order structure represents an estimated common structure most indicative of the macromolecules under investigation, while allowing the possibility of probabilistic variation. We do not, however, conclude that the consensus structure be applied to a specific single sequence

directly due to the possibility of alignment error. Some small adjustments may be required when applied to individual sequences.

By satisfying both the statistical and structural constraints in our method, the resulting consensus structure of the macromolecules has position pairings with the strongest statistical interdependency. This evaluation can be applied irrespective of the residue type. Substructures of the macromolecules (e.g. helices) are identified by position pairing characteristics. Structural planarity property of the pairings is used to differentiate secondary pairings from tertiary pairings. The method is based on the idea that a folding pattern of the macromolecules must demonstrate a strong statistical interdependence between positions of direct interactions. The type of bonding between these interactive positions is not necessarily restricted to the Watson–Crick base pairing.

We have evaluated the method on several ensembles of test data, including an ensemble of aligned data of tRNA and 5S rRNA sequences. Compared with the well known structure of *yeast tRNA^{Pha}* and *Escherichia coli 5S rRNA*, the generated higher-order structure in our analysis conforms well to the reported structures.

The method can be extended to detect other types of substructures in macromolecules, for example, the detection of pseudoknots and protein sequences (Pleij and Bosch, 1989). The algorithmic complexity of the method is polynomial and the computational time required in the calculation is small. Currently, experimental studies have been carried out on 16S rRNA sequences and protein sequences and will be reported separately.

Acknowledgements

The authors would like to thank Dr George Harauz for helpful discussions. The research was supported by Natural Sciences and Engineering Research Council of Canada Operating Grant 865-42 and a grant from the University of Guelph Research Enhancement Fund.

References

Ash,R. (1965) *Information Theory*, Wiley-Interscience, New York

Benedetti,G., De Santis,P. and Morosetti,S. (1989) A new method to find a set of energetically optimal RNA secondary structures. *Nucleic Acids Res.*, **13**, 5149–5161.

Chan,S C., Wong,A K C and Chiu,D.K.Y. (1991) A survey of multiple sequence comparison methods. *Bull. Math. Biol.*, in press.

Chiu,D.K.Y., Cheung,B. and Wong,A.K.C. (1990) Information synthesis based on hierarchical maximum entropy discretization. *J. Exp. Theor. Artif. Intell.*, **2**, 117–129

Chiu,D.K.Y , Wong,A.K.C. and Cheung,B. (1991) Information discovery through hierarchical maximum entropy discretization and synthesis. In *Knowledge Discovery in Databases*, MIT Press, in press.

Garnier,J., Osguthorpe,D.J. and Robson,B. (1978) Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins. *J. Mol. Biol*, **120**, 97–120.

Gibrat,J -F., Garnier,J. and Robson,B. (1987) Further developments of protein secondary structure prediction using information theory, new parameters and consideration of residue pairs. *J. Mol. Biol.*, **198**, 425–443.

Gouy,M. (1987) Secondary structure prediction of RNA. In Bishop,M J and Rawlings,C.J. (eds) *Nucleic Acid and Protein Sequence Analysis: A Practical Approach*, IRL Press, Oxford, pp. 259–284.

Grosjean,H., Cedergren,R.J. and McKay,W. (1982) Structure in tRNA data. *Biochimie*, **64**, 387–397.

Haselman,T., Camp,D G. and Fox,G.E (1989) Phylogenetic evidence for tertiary interactions in 16S-like ribosomal RNA. *Nucleic Acids Res* , **17**, 2215–2221.

Haselman,T , Chappelear,J E. and Fox,G.E. (1988) Fidelity of secondary and tertiary interactions in tRNA. *Nucleic Acids Res.*, **16**, 5673–5684.

Jaeger,J A., Turner,D.H., and Zuker,M. (1990) Predicting optimal and sub-optimal secondary structure for RNA. *Methods Enzymol.*, **183**, 281–306.

Mortimer,E. (1979) Review of Blaise Pascal: The Life and Work of a Realist. *Mathamatics, An Introduction to Its Spirits and Use, Readings from Scientific American*, W.H.Freeman, San Francisco.

Nussinov,R , Pieczenik,G., Griggs,J.R and Kleitmans,D.J. (1978) Algorithms for loop matchings *SIAM J. Appl Math.*, **35**, 68–82.

Olsen,G.J. (1988) Phylogenetic analysis using ribosomal RNA. *Methods Enzymol* , **164**, 793–812.

Pleij,C.W.A. and Bosch,L. (1989) RNA pseudoknots: structure, detection, and prediction. *Methods Enzymol.*, **180**, 289–303.

Sussman,J.L. and Kim,S. (1976) Three-dimensional structure of a transfer RNA in two crystal forms. *Science*, **192**, 853–858.

Tinoco,I., Uhlenbeck,O.C. and Levine,M.D. (1971) Estimation of secondary structure in ribonucleic acids. *Nature*, **230**, 362–367

Waterman,M.S. (1984) General methods of sequence comparison *Bull. Math Biol.*, **46**, 473–500.

Waterman,M.S. (1986) Multiple sequence alignment by consensus. *Nucleic Acids Res.*, **14**, 9095–9102.

Waterman,M S. (1988) Computer analysis of nucleic acid sequences. *Methods Enzymol.*, **164**, 765–793.

Waterman,M.S. and Smith,T.F. (1978) RNA secondary structure: a complete mathematical analysis. *Math. Bios.*, **42**, 257–266.

Wong,A.K.C. and Chiu,D.K.Y (1987) An event covering method for effective probabilistic inference. *Pattern Recognition*, **20**, 245–255.

Wong,A.K C. and Liu,T S. (1975) Typicality, diversity and feature pattern of an ensemble. *IEEE Trans. Comput* , **24**, 158–181

Zuker,M. and Sankoff,D. (1984) RNA secondary structures and their prediction. *Bull. Math. Biol.*, **46**, 591–621

Zuker,M. and Stiegler,P. (1981) Optimal computer folding of large RNA sequences using thermodynamics and auxilliary. *Nucleic Acids Res.*, **9**, 133–148

Received on July 20, 1990, accepted on January 15, 1991