BIOINFORMATICS

Vol. 19 no. 5 2003, pages 587–592
DOI: 10.1093/bioinformatics/btg032

The fuzzy polynucleotide space: basic properties

Angela Torres[1] *and Juan J. Nieto*[2,*]

[1]*Departamento de Psiquiatría, Radiología y Salud Pública, Facultad de Medicina, Universidad de Santiago de Compostela, Spain and* [2]*Departamento de Análisis Matemático, Facultad de Matemáticas, Universidad de Santiago de Compostela, Spain*

Received on June 5, 2002; revised on September 25, 2002; accepted on November 4, 2002

ABSTRACT

Motivation: Any triplet codon may be regarded as a 12-dimensional fuzzy code. Sufficient information about a particular sequence may not be available in certain situations. The investigator will be confronted with imprecise sequences, yet want to make comparisons of sequences. Fuzzy polynucleotides can be compared by using geometrical interpretation of fuzzy sets as points in a hypercube.

Results: We introduce the space of fuzzy polynucleotides and a means of measuring dissimilitudes between them. We establish mathematical principles to measure dissimilarities between fuzzy polynucleotides and present several examples in this metric space.

We calculate the frequencies of the nucleotides at the three base sites of a codon in the coding sequences of *Escherichia coli K-12* and *Mycobacterium tuberculosis H37Rv*, and consider them as points in that fuzzy space. We compute the distance between the genomes of *E.coli* and *M.tuberculosis*.

Availability: Available on request from the authors.

Contact: amnieto@usc.es

1 MOTIVATION

The Human Genome Project has been regarded as a major enterprise of science in the 20th century. The genetic material in our chromosomes is our genome. Chromosomes are long DNA molecules. The structure of DNA was discovered by Watson and Crick in 1953. The authors showed that a DNA molecule is a double helix consisting of two strands. Each helix is a chain of *bases*, chemical units of four types: T, C, A, and G.

The hereditary instructions are written in this four-letter alphabet, each letter corresponding to one of the chemical constituents of DNA (nitrogen-containing chemicals termed bases): T (thymine), C (cytosine), A (adenine), and G (guanine). The sequence of Ts, Cs, As, and Gs are the recipe for a specific protein. Twenty different amino acids constitute the building blocks of all proteins. Thus, the genetic information is encoded digitally, as strings over the four-letter alphabet, {T, C, A, G}, much as information is encoded digitally in computers as strings of zeros and ones (Karp, 2002). The DNA alphabet is {T, C, A, G} , and the RNA alphabet, {U, C, A , G}, U (uracil).

Computer Science, Statistics, Artificial Intelligence, and Mathematics help in the management and analysis of the Human Genome Project's massive database. For most of the twentieth century, mathematics played a minor role in biology and genomics, but its role has expanded greatly over the last twenty years (Speed, 2002). The fusion of mathematics and biology will result in a new era of molecular medicine, when the diagnosis, treatment, and prevention of disease will be individually oriented and, therefore, more successful. For general mathematical aspects of the human genome and genetic analysis we refer to Ewens (2001), Karp (2002), Lange (2002), Paun *et al.* (1998), Percus (2002) and Speed (2002). Further mathematical contributions, to mention a few, are the following: A simple mathematical model to evaluate the efficiency and feasibility of two strategies in DNA cloning (Tang, 2000), discrepancy of DNA sequences (Fang and Roberts, 2001), distance on DNA knots (Darcy, 2001), algorithms for the alignment of DNA sequences (Lee *et al.*, 2002; Lenhof *et al.*, 1999; Morgenstern, 2002), sequence similarities in genomes (Vicens *et al.*, 2001), sequence distances (Foster *et al.*, 1999; Li *et al.*, 2001; Liben-Nowell, 2001), or dissimilarity of DNA sequences (Wu *et al.*, 1997). Further publications in which mathematics plays a central role are the following: (Alves and Savageau, 2000; Demetrius *et al.*, 1985; Ferreira *et al.*, 2002; Goryanin *et al.*, 1999; Jamshidi *et al.*, 2001; Kargupta, 2001; Lee *et al.*, 2002; Pevzner *et al.*, 2001).

In Sadegh-Zadeh (2000), the author showed that a polynucleotide can be represented as an ordered fuzzy set. The genetic code may be considered to be 12 dimensional, as a triplet codon XYZ has a $3 \times 4 = 12$-dimensional fuzzy code $(a_1, a_2, \ldots, a_{12})$. Thus, it is a point in the 12-dimensional fuzzy polynucleotide space I^{12} with $I = [0, 1] \subset \mathbf{R}$.

*To whom correspondence should be addressed.

Fuzzy logic and fuzzy technology is now frequently used in bioinformatics. For example, fuzzy logic is used to increase the flexibility of protein motifs (Chang and Halgamuge, 2002), and fuzzy adaptive resonance theory is utilized to analyze experimental expression data (Tomida *et al.*, 2002). A novel approach to local reliability of sequence alignments based on a fuzzy recast of the dynamic programming algorithm is presented in Schlosshauer and Ohlsson (2002).

2 THE FUZZY HYPERCUBE

Kosko (1992) introduced a geometrical interpretation of fuzzy sets as points in a hypercube. Indeed, for a given set

$$X = \{x_1, \ldots, x_n\}$$

a fuzzy subset is just a mapping

$$\mu : X \to I = [0, 1],$$

and the value $\mu(x)$ expresses the grade of membership of the element $x \in X$ to the fuzzy subset μ. Thus, the set of all fuzzy subsets (of X) is precisely the unit hypercube

$$I^n = [0, 1]^n,$$

as any fuzzy subset μ determines a point

$$P \in I^n$$

given by

$$P = (\mu(x_1), \ldots, \mu(x_n)).$$

Reciprocally, any point $A = (a_1, \ldots, a_n) \in I^n$ generates a fuzzy subset μ defined by $\mu(x_i) = a_i$, $i = 1, \ldots, n$.

Nonfuzzy or crisp subsets of X are given by mappings

$$\mu : X \to \{0, 1\},$$

and are located at the 2^n corners of the n-dimensional unit hypercube I^n. We have represented the fuzzy hypercube I^2 in Figure 1, and I^3 in Figure 2. The n-dimensional hypercube I^n is graphically not representable for $n \geq 4$.

Fuzzy logic plays an important role in medicine (Abbod *et al.*, 2001; Barro and Marín, 2002; Mahfouf *et al.*, 2001; Nieto and Torres, 2002; Pickert *et al.*, 1998; Sadegh-Zadeh, 2000; Szczepaniak *et al.*, 2000), and in bioinformatics (Chang and Halgamuge, 2002; Dougherty *et al.*, 2002; Pickert *et al.*, 1998; Schlosshauer and Ohlsson, 2002; Tomida *et al.*, 2002).

Hypercubical calculus has been described in Zaus (1999), while some biomedical applications of the fuzzy unit hypercube are given in Helgason and Jobe (1998); Nieto and Torres (2002) and Sadegh-Zadeh (1999).

Any codon corresponds to a corner of the 12-dimensional unit hypercube I^{12}. Any element of I^{12} may be viewed as a fuzzy codon. Therefore, we consider I^{12} as the set of fuzzy polynucleotides.

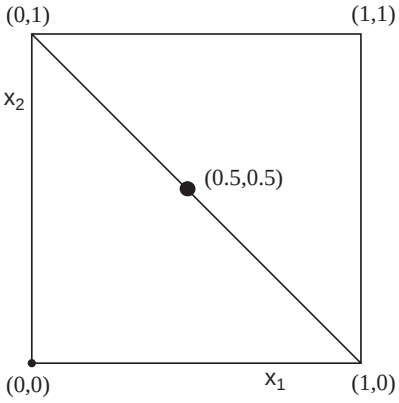

Fig. 1. Two-dimensional hypercube I^2 with the 4 nonfuzzy subsets (0,0), (1,0), (0,1), (1,1), and the fuzzy set (0.5,0.5).

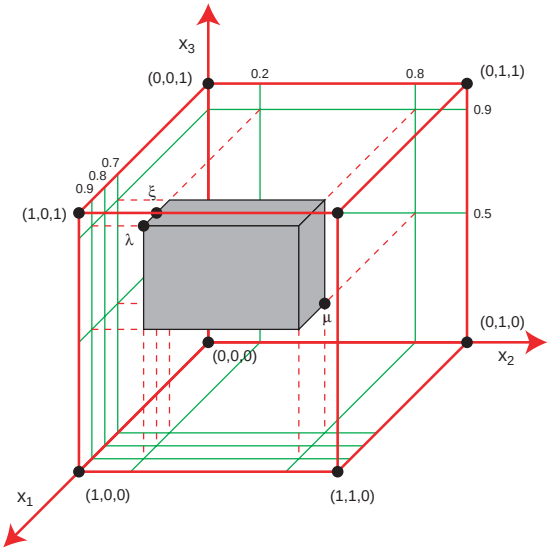

Fig. 2. Three-dimensional hypercube I^3 with its $2^3 = 8$ nonfuzzy sets. The points $\lambda = (0.9, 0.2, 0.9)$, $\mu = (0.7, 0.8, 0.5)$, and $\xi = (0.8, 0.2, 0.9)$ represent fuzzy sets.

For a complete genome, we consider the frequencies of the nucleotides at the three base sites of a codon in the coding sequence. It may be viewed as a point in the hypercube I^{12}. We apply this idea to the genomes of *M. tuberculosis* and *E. coli* to obtain their fuzzy set of frequencies.

3 THE FUZZY SET OF POLYNUCLEOTIDES

The DNA alphabet is { T, C, A , G }, and the RNA alphabet { U, C, A , G }. Consider the RNA alphabet. If U, the first

letter of this alphabet, appears, we code it as 1 0 0 0 :

- 1 since the first letter (U) is present;
- 0 since the second letter (C) does not appear;
- 0 since the third letter (A) is not present; and
- 0 since the fourth letter (G) does not appear.

Thus, C is coded as 0 1 0 0, A is coded as 0 0 1 0, and G as 0 0 0 1.

The genetic code may be considered to be 12 dimensional, as a triplet codon XYZ has a $3 \times 4 = 12$-dimensional fuzzy code $(a_1, a_2, \ldots, a_{12})$, a point in the 12-dimensional fuzzy polynucleotide space I^{12}.

Any of the 64 codons of the genetic code is located at one of the $2^{12} = 4\,096$ corners of this 12-dimensional unit hypercube I^{12}. For example, the codon CAU corresponds to the amino acid histidine; its fuzzy code is then

$$(\underbrace{0, 1, 0, 0}_{C} , \underbrace{0, 0, 1, 0}_{A} , \underbrace{1, 0, 0, 0}_{U}) \in I^{12}.$$

However, sufficient knowledge of the chemical structure of a particular sequence may not be available in all cases. We may be dealing with base sequences that are not necessarily at a corner of the hypercube, with the components of its fuzzy code being neither 0 nor 1. For example,

$$(0.3, 0.4, 0.1, 0.2, 0, 1, 0, 0, 0, 0, 0, 1) \in I^{12}$$

represents a codon XCG, i.e. C in the second position, G in the third position, and the first letter is

- U to the extent 0.3,
- C to the extent 0.4,
- A to the extent 0.1,
- G to the extent 0.2.

If the first letter would be C, we would have the codon CCG (proline); if the first letter is U, we would have the codon UCG (serine). If we have to select one, we would probably choose CCG, as $0.4 > 0.3$.

This is a typical situation in *sequence analysis* where some information is missing. This would correspond, for example, to emission probabilities in hidden Markov models (Husmeier and Wright, 2001). Comparison of sequences, however, is a fundamental task. We then have to determine distances, differences, similarities, and dissimilarities between chains of nucleic acid (Darcy, 2001; Fang and Roberts, 2001; Wu *et al.*, 1997) and, consequently, between polynucleotides.

We introduce a distance, a *metric* in mathematical terminology, in the fuzzy unit hypercube I^{12} to obtain the Fuzzy Polynucleotide Space. Of course, we could analyze sequences of any length by considering $I^{12\,k}$ with k being the length of the sequence. For example, the sequence UACUGU (tyrosine/cysteine) is a point in I^{24}.

Computing the frequencies of the nucleotides at the three base sites for a complete genome sequence, we obtain a point in the fuzzy unit hypercube I^{12}.

4 THE FUZZY POLYNUCLEOTIDE SPACE

Consider the 12-dimensional unit hypercube. To measure how different two fuzzy polynucleotides are, we introduce a distance between them. Given,

$$p = (p_1, p_2, \ldots, p_n),\ q = (q_1, q_2, \ldots, q_n) \in I^n,\ n = 12,$$

not both equal to the empty set $\emptyset = (0, 0, \ldots, 0)$, we define the difference between p and q as

$$d(p, q) = \frac{\sum_{i=1}^{12} |p_i - q_i|}{\sum_{i=1}^{12} \max\{ p_i , q_i \}}. \qquad (1)$$

Of course, $d(\emptyset, \emptyset) = 0$.

The distance defined by Equation (1) is motivated by the publications Lin (1997) and Sadegh-Zadeh (2000).

We know that d is indeed a metric (Nieto *et al.*, 2003), as it satisfies the properties of non-negativity, symmetry, and the triangle inequality. Thus (I^{12}, d) is a metric space, and we obtain the Fuzzy Polynucleotide Metric Space.

A few examples:

$$d(\text{histidine, proline}) = d(\text{CAU, CCG}) = \frac{4}{5} = 0.8,$$

$$d(\text{histidine, serine}) = d(\text{CAU, UCG}) = 1,$$

$$d(\text{arginine, glutamine}) = d(\text{CGU, CAG}) = \frac{4}{5} = 0.8,$$

$$d(\text{histidine, arginine}) = d(\text{CAU, CGU}) = \frac{1}{2} = 0.5,$$

$$d(\text{lysine, glycine}) = d(\text{AAA, GGG}) = 1.$$

We should note that the genetic code is degenerated, e.g. histidine can be represented by two different codons (CAU and CAC). We distinguish between both codons. We may fuse, in the future, some corners of the unit hypercube.

Now, let XCG$=$ $(0.3, 0.4, 0.1, 0.2, 0, 1, 0, 0, 0, 0, 0, 1) \in I^{12}$. It is evident that the codon XCG is closer to proline than to serine as is indicated by the following distances

$$d(\text{XCG, proline}) = d(\text{XCG, CCG}) = \frac{1}{3} \approx 0.3333,$$

$$d(\text{XCG, serine}) = d(\text{XCG, UCG}) = \frac{1.4}{3.7} \approx 0.3784.$$

The program to evaluate these distances is very simple; it is available on request from the authors.

5 A STRIKING PROPERTY OF THE FPS

Based on our knowledge of Euclidean geometry, we say that a point R is between two given points P and Q if

$$d(P, Q) = d(P, R) + d(Q, R).$$

In the usual Euclidean space, the set of points between P and Q is precisely the set of points of the Euclidean segment joining P and Q.

However, given two points of the fuzzy polynucleotide space, we show that there is no point between them. Indeed, let $P = (0, 0, \ldots, 0) = \emptyset$, and $Q \in I^{12}$, $Q \neq \emptyset$. Then, there is no point between P and Q, i.e. if $R \in I^{12}$ we have

$$d(P, Q) < d(P, R) + d(R, Q),$$

since $d(\emptyset, R) = 1$, $d(\emptyset, R) = 1$, and $d(R, Q) > 0$.

The point $P = \emptyset$ corresponds to the empty set (*no polynucleotide* $\equiv$ *no information*) which is entirely different from any fuzzy polynucleotide, since $d(\emptyset, Q) = 1$ for any $Q \in I^{12}$. Thus, *information/life* and *no information/no life* are two separate entities with no intervening or connecting element.

6 APPLICATION: DIFFERENCES BETWEEN GENOMES

Statistical and fuzzy analysis of DNA sequences can provide useful insight of the structures of genomes (Chen *et al.*, 2002; de Sousa, 1999; Nyeo *et al.*, 2002). The frequencies of the nucleotides at the three base sites of a codon in the coding sequence of a genome may be viewed as a point in I^{12}. We calculate those frequencies for two bacteria.

The complete genome sequence of *Mycobacterium tuberculosis H37Rv* is available at http://www.ncbi.nlm.nih.gov. Its accession number is NC_000962. The genome comprises 4 411 529 base pairs, contains around 4000 genes, and has a very high guanine+cytosine content (Cole *et al.*, 1998).

We compute the number of the nucleotides at the three base sites of a codon in the coding sequences of *M.tuberculosis* (see Table 1), and then calculate the corresponding fractions (Table 2). We define, as usual, the noncoding sequences as the common intersecting regions of the noncoding sequences on both strands. For the analysis of noncoding sequences, we consider the sequences if they have a length of 512 base pairs or longer (Nyeo *et al.*, 2002). For example, at the first base we have 216 051 T of a total of 1 324 174. Thus, the fraction of T in the first base is

$$\frac{2\,160\,510}{1\,324\,174} = 0,1632 = 16.32\%.$$

Table 1. Number of nucleotides at the three base sites of a codon in the coding sequence of *Mycobacterium tuberculosis*

	T	C	A	G
First base	216 051	409 011	228 244	470 868
Second base	269 638	416 457	233 472	404 607
Third base	217 803	458 256	210 892	437 223

Table 2. Fractions of nucleotides at the three base sites of a codon in the coding sequence of *Mycobacterium tuberculosis*

	T	C	A	G
First base	0.1632	0.3089	0.1724	0.3556
Second base	0.2036	0.3145	0.1763	0.3056
Third base	0.1645	0.3461	0.1593	0.3302

Table 3. Number of nucleotides at the three base sites of a codon in the coding sequence of *Escherichia coli*

	T	C	A	G
First base	215 406	324 793	348 972	452 813
Second base	418 197	306 729	381 930	235 128
Third base	351 502	344 638	245 774	400 070

Table 4. Fractions of nucleotides at the three base sites of a codon in the coding sequence of *Escherichia coli*

	T	C	A	G
First base	0.1605	0.2420	0.2600	0.3374
Second base	0.3116	0.2286	0.2846	0.1752
Third base	0.2619	0.2568	0.1831	0.2981

We may consider these fractions as a point in the hypercube I^{12}. Indeed, the point

$$(0.1632, 0.3089, 0.1724, 0.3556, 0.2036, 0.3145, 0.1763, 0.3056, 0.1645, 0.3461, 0.1593, 0.3302) \in I^{12}.$$

This is the fuzzy set of frequencies of the genome sequence of *M.tuberculosis*.

The 4 639 221 base pair sequence of *Escherichia coli K-12* is presented in Blattner *et al.* (1997). It is also available at http://www.ncbi.nlm.nih.gov, with accession number NC_000913. The number and frequencies of the nucleotides at the three base sites of a codon in the coding sequences of *E.coli* are presented in Tables 3 and 4, respectively. The fuzzy set of frequencies of the genome

of *E.coli* is

$$(0.1605, 0.2420, 0.2600, 0.3374, 0.3116, 0.2286, 0.2846, 0.1752, 0.2619, 0.2568, 0.1831, 0.2981) \in I^{12}.$$

Using the distance given in Equation (1), we compute the distance between these two fuzzy sets representing the frequencies of the nucleotides of *M.tuberculosis* and *E.coli*:

$$d(M.tuberculosis, E.coli) = \frac{0.8506}{3.4253} \approx 0.2483.$$

7 CONCLUSIONS AND FURTHER RESEARCH

We introduce the Fuzzy Polynucleotide Space to compare fuzzy codons. A representative example is presented to compare complete genomes.

This is a basic and theoretical study, but we hope that a deeper development of our ideas will help to find some relevant information in genomes. For example, one may use different alphabets (the two-letters alphabet of purine–pyrimidine, the twenty-letters alphabet of proteins,. . . .), or one may be interested in how to weight the first two and the last position in a codon since the last of the three codes in a code group is wobbly (Garner *et al.*, 1998).

The metric given by Equation (1) may have many undiscovered yet properties. A further study of the Fuzzy Polynucleotide Space is necessary, and then to interpret those properties biologically.

The average composition of a sequence could be useful to describe motifs.

ACKNOWLEDGEMENTS

Research partially supported by Ministerio de Ciencia y Tecnología and FEDER, project BFM2001–3884–C02–01, and by Xunta de Galicia and FEDER, project PGIDIT02PXIC20703PN. The authors thank the anonymous referees for their useful remarks and interesting suggestions.

REFERENCES

Abbod,M.F., von Keyserlingk,D.G., Linkens,D.A. and Mahfouf,M. (2001) Survey of utilisation of fuzzy technology in medicine and healthcare. *Fuzzy Sets and Systems*, **120**, 331–349.

Adleman,L. (1994) Molecular computation of solutions to combinatorial problems. *Science*, **226**, 1021–1024.

Alves,R. and Savageau,M.A. (2000) Extending the method of mathematically controlled comparison to include numerical comparisons. *Bioinformatics*, **16**, 786–798.

Barro,S. and Marín,R. (2002) *Fuzzy Logic in Medicine*. Physica, Heidelberg.

Blattner,F.R., Plunkett,G., Bloch,C.A. *et al.*, (1997) The complete genome sequence of *Escherichia coli K-12*. *Science*, **277**, 1453–1474.

Chang,B.C.H. and Halgamuge,S.K. (2002) Protein motif extraction with neuro-fuzzy optimization. *Bioinformatics*, **18**, 1084–1090.

Chen,Y., Kamat,V., Dougherty,E.R., Bittner,L., Meltzer,P.S. and Trent,M. (2002) Ratio statistics of gene expression levels and applications to microarray data analysis. *Bioinformatics*, **18**, 1207–1215.

Cole,S.T., Brosch,R., Parkhill,J. *et al.*, (1998) Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. *Nature*, **393**, 537–544.

Darcy,I.K. (2001) Biological distances on DNA knots and links: applications to XER recombination. *Journal of Knot Theory and Ramifications*, **10**, 269–294.

de Sousa,M. (1999) Statistics of DNA sequences: a low-frequency analysis. *Phys. Rev. E*, **60**, 5932–5937.

Demetrius,L., Schuster,P. and Sigmund,K. (1985) Polynucleotide evolution and branching processes. *Bull. Math. Biol.*, **47**, 239–262.

Dougherty,E.R., Barrera,J., Brun,M., Kim,S., Cesar,R.M., Chen,Y., Bittner,M. and Trent,J.M. (2002) Inference from clustering with application to gene-expression microarrays. *J. Comput. Biol.*, **9**, 105–126.

Ewens,W.J. (2001) Mathematics and the human genome project. *The Mathematical Scientist*, **26**, 1–12.

Fang,W. and Roberts,F.S. (2001) A measure of discrepancy of multiple sequences. *Information Sciences*, **137**, 75–102.

Ferreira,C.E., de Souza,C.C. and Wakabayashi,Y. (2002) Rearrangement of DNA fragments: a branch-and-cut algorithm. *Discrete Applied Mathematics*, **116**, 161–177.

Foster,M., Heath,A. and Afzal,M. (1999) Application of distance geometry to 3D visualization of sequence relationships. *Bioinformatics*, **15**, 89–90.

Garner,E., Cannon,P., Romero,P., Obradovic,Z. and Dunker,A.K. (1998) Predicting disordered regions from amino acid sequence: common themes despite differing structural characterization. *Genome Informatics*, **9**, 201–214.

Goryanin,I., Hodgman,T.C. and Selkov,E. (1999) Mathematical simulation and analysis of cellular metabolism and regulation. *Bioinformatics*, **15**, 749–758.

Helgason,C.M. and Jobe,T.H. (1998) The fuzzy cube and causal efficacy: representation of concomitant mechanisms in stroke. *Neural Networks*, **11**, 549–555.

Husmeier,D. and Wright,F. (2001) Detection of recombination in DNA multiple alignment with hidden Markov models. *J. Comput. Biol.*, **8**, 401–427.

Jamshidi,N., Edwards,J.S., Fahland,T., Church,G.M. and Palsson,B.O. (2001) Dynamic simulation of the human red blood cell metabolic network. *Bioinformatics*, **17**, 286–287.

Kargupta,H. (2001) A striking property of genetic code-like transformations. *Complex Systems*, **13**, 1–32.

Karp,R.M. (2002) Mathematical challenges from genomics and molecular biology. *Notices of the American Mathematical Society*, **49**, 544–553.

Kosko,B. (1992) *Neural networks and fuzzy systems*. Prentice-Hall, Englewood Cliffs, NJ.

Lange,K. (2002) *Mathematical and statistical methods for genetic analysis*. Springer, New York.

Lee,C., Grasso,C. and Sharlow,M.F. (2002) Multiple sequence alignment using partial order graphs. *Bioinformatics*, **18**, 452–464.

Lenhof,H.P., Morgenstern,B. and Reinert,K. (1999) An exact solution for the segment-to-segment multiple sequence alignment problem. *Bioinformatics*, **15**, 203–210.

Li,M., Badger,J.H., Chen,X., Kwong,S., Kearney,P. and Zhang,H. (2001) An information-based sequence distance and its application to whole mitochondrial phylogeny. *Bioinformatics*, **17**, 149–154.

Liben-Nowell,D. (2001) On the structure of syntenic distance. *J. Comput. Biol.*, **8**, 53–67.

Lin,C.T. (1997) Adaptive subsethood for radial basis fuzzy systems. In Kosko,B. (ed.), *Fuzzy Engineering*. Prentice-Hall, Upper Saddle River, NJ, pp. 429–464.

Mahfouf,M., Abbod,M.F. and Linkens,D.A. (2001) A survey of fuzzy logic monitoring and control utilisation in medicine. *Artificial Intelligence in Medicine*, **21**, 27–42.

Morgenstern,B. (2002) A simple and space-efficient fragment-chaining algorithm for alignment of DNA and protein sequences. *Appl. Math. Lett.*, **15**, 11–16.

Nieto,J.J. and Torres,A. (2002) Midpoints for fuzzy sets and their application in medicine. *Artificial Intelligence in Medicine*, in press.

Nieto,J.J., Torres,A. and Vázquez-Trasande,M.M. (2003) A metric space to study differences between polynucleotides. *Appl. Math. Lett.*, **27**, 81–101.

Nyeo,S.-L., Yang,I.-C. and Wu,C.-H. (2002) Spectral classification of archaeal and bacterial genomes. *J. Biol. Syst.*, **10**, 233–241.

Paun,Gh., Rozenberg,G. and Salomaa,A. (1998) *DNA Computing: New Computing Paradigms*. Springer, Berlin.

Percus,J. (2002) *Mathematics of Genome Analysis*. Cambridge University Press, Cambridge.

Pevzner,P.A., Tang,H. and Waterman,M.S. (2001) An Eulerian path approach to DNA fragment alignment. *Proc. Natl Acad. Sci. USA*, **98**, 9748–9753.

Pickert,I., Reuter,I., Klawonn,F. and Wingender,E. (1998) Transcription regulatory region analysis using signal detection and fuzzy clustering. *Bioinformatics*, **14**, 244–251.

Sadegh-Zadeh,K. (1999) Fundamentals of Clinical Methodology: 3. Nosology. *Artificial Intelligence in Medicine*, **17**, 87–108.

Sadegh-Zadeh,K. (2000) Fuzzy genomes. *Artificial Intelligence in Medicine*, **18**, 1–28.

Sadegh-Zadeh,K. (2001) The fuzzy revolution: goodbye to the Aristotelian weltanschauung. *Artificial Intelligence in Medicine*, **21**, 1–25.

Schlosshauer,M. and Ohlsson, (2002) A novel approach to local reliability of sequence alignments. *Bioinformatics*, **18**, 847–854.

Speed,T. (2002) Mathematics and the human genome. *Notices of the American Mathematical Society*, **49**, 429.

Szczepaniak,P.S., Lisboa,P.J.G. and Kacprzyk,J. (2000) *Fuzzy Systems in Medicine*. Physica, Heidelberg.

Tang,B. (2000) Evaluation of some DNA cloning strategies. *Computers and Mathematics with Applications*, **39**, 43–48.

Tomida,S., Hanai,T., Honda,H. and Kobayashi,T. (2002) Analysis of expression profile using fuzzy adaptive resonance theory. *Bioinformatics*, **18**, 1073–1083.

Vicens,P., Badel-Chagnon,A., André,C. and Hazout,S. (2001) D-ASSIRC: distributed program for finding sequence similarities in genomes. *Information Sciences*, **137**, 75–102.

Wu,T.J., Burke,J.P. and Davison,D.B. (1997) A measure of DNA sequence dissimilarity based on Mahalonobis distance between frequencies of words. *Biometrics*, **53**, 1431–1439.

Zaus,M. (1999) *Crisp and Soft Computing with Hypercubical Calculus*. Physica, Heidelberg.